Journal of Experimental Psychology: General
2015, Vol. 144, No. 1, 76–84

0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000030

Finding a Needle in a Haystack: Toward a Psychologically Informed Method for Aviation Security Screening

Thomas C. Ormerod
University of Sussex

Coral J. Dando
University of Wolverhampton

Current aviation security systems identify behavioral indicators of deception to assess risks to flights, but they lack a strong psychological basis or empirical validation. We present a new method that tests the veracity of passenger accounts. In an in vivo double-blind randomized-control trial conducted in international airports, security agents detected 66% of deceptive passengers using the veracity test method compared with less than 5% using behavioral indicator recognition. As well as revealing advantages of veracity testing over behavioral indicator identification, the study provides the highest levels to date of deception detection in a realistic setting where the known base rate of deceptive individuals is low.

You have to measure whether what we're doing is the only way to assure . . . safety. And you also have to think are there ways . . . that are less intrusive (Barak Obama, Lisbon, November 2010).

Identifying threats presents a huge challenge to those tasked with ensuring public safety, and to psychologists developing methods for detecting deception. However, the news from both arenas is not good. Since the events of September 11, 2001, billions of dollars have been invested in aviation security procedures designed to detect threats to airplanes (United States Government Accountability Office, 2011), but the effectiveness of these procedures has been questioned (Weinberger, 2010). More recent events such as the 2009 attempted bombing of Flight NW253 to Detroit suggest we still lack effective ways of identifying threats to public safety.

Threat detection procedures typically involve looking for individuals who display behaviors thought to be indicators of deception, particularly behaviors shown by the perpetrators of previous attacks. The effectiveness of behavioral indicator approaches has never been tested in a large-scale field trial. However, a meta-analysis of laboratory studies that used behavioral indicators to discriminate deceivers from truth-tellers revealed a mean rate for correct identification of only 54% (Bond & DePaulo, 2006). This rate was significantly greater than 50%, the modal percentage of truth-tellers in samples included in their analysis, but it inspires little confidence in the prospects for detecting a deceiver among many hundreds of truth-tellers, as would likely be the case in aviation security screening. According to Levine (2010), the slight but above significance rate of discriminating truth-tellers from deceivers arises, not because judges of deceptive behavior have some degree of competence at identifying relevant behavioral indicators, but because there are generally a few deceivers in any study who are particularly poor at masking their lies.

In this article, we present a new procedure for aviation security screening that is based, not on behavioral indicators selected from previous incidents, but on testing the veracity of passengers' verbal accounts. Our approach takes techniques derived from psychological theory and shown in recent laboratory studies to yield promising rates of deception detection, and integrates them into a comprehensive procedure for detecting threat. We then compare the effectiveness of behavioral-indicator and veracity-testing approaches in an in vivo empirical evaluation conducted with passengers departing on flights at international airports during routine security screening.

Approaches to Detecting Deception

The majority of published research on detecting deception has set out to identify indicators in human behavior that can discriminate deceivers from truth-tellers. Behavioral indicators of deception fall into two main categories: physical behaviors relating to demeanor (e.g., nervousness, aggression) and/or actions (e.g., eye contact, fidgeting); and verbal behaviors relating to the nature and production of speech (e.g., hesitations, use of pronouns). As noted above, behavioral indicator approaches typically yield low discrimination rates. Indeed, Bond and DePaulo (2006) found accuracy at discriminating truth-tellers from deceivers was lower when judgments were made from visual rather than auditory media. They suggest that judges derive judgments by misusing nonverbal cues to affective states (e.g., guilt, anxiety, or shame) that are stereotypically associated with lying. There is some evidence that

This article was published Online First November 3, 2014.

Thomas C. Ormerod, School of Psychology, University of Sussex; Coral J. Dando, Institute of Psychology, University of Wolverhampton.

The data reported in this article are available from the authors. The research was funded by Her Majesty's Government Communications Centre, United Kingdom (HMGCC). A detailed protocol of the study can be obtained from the authors. The authors gratefully acknowledge the assistance of Alexandra Sandham and Jane Stuart who acted as independent coders for interrater reliability checks and subsequently assisted in data collection, Padraic Monaghan, Paul Taylor, and Mark Howe who commented on earlier drafts, and Henry Roediger III and two anonymous reviewers for suggested improvements to the manuscript.

Correspondence concerning this article should be addressed to Thomas C. Ormerod, School of Psychology, University of Sussex, Falmer, Sussex, BN1 9QH, UK. E-mail: t.ormerod@sussex.ac.uk

counts of illustrators (i.e., hand movements to indicate content or prosody) can provide reasonable levels of discrimination (DePaulo et al., 2003), but the range of practical contexts in which illustrators can be used is limited (e.g., real-time detection of differences in the use of illustrators is likely to prove impractical).

Low rates of deception detection from behavioral indicators arise, according to Levine, Kim, and Blair (2010), for four reasons: a lack of indictors with predictive validity; naive beliefs in the predictive validity of certain indicators (e.g., avoidance of eye contact); ignoring information that may indicate deception (e.g., failing to spot inconsistencies in an account); and truth bias, that is, a predisposition to assume the truth of another person's account. The Situational Familiarity theory of Stiff et al. (1989) posits that individuals will place greater credence on their ability to judge the verbal accounts of others when they describe familiar situations (e.g., events occurring in locations known to them). Consequently, they are more likely to judge accounts about familiar situations on the basis of verbal content, but are more likely to rely on nonverbal cues when the situation is unfamiliar. Reinhard, Sporer, Scharmach, and Marksteiner (2011) found effects of situational familiarity (both actual and perceived) on deception detection, with greater judgment accuracy when the situation was familiar, an effect they demonstrated to be because of greater reliance on verbal content with familiar situations.

In their "Dangerous Decisions Theory," Porter, Gustaw, and ten Brinke (2010) argue that an initial schema formed from misinterpreted behavioral indicators biases people's judgments about deception, leading to irrational decision-making in the face of contradictory evidence. Despite these concerns, it has been suggested that deception research has "been characterized by a myopic focus on the internal psychological states and corresponding nonverbal behaviors of liars and has failed to adequately consider the situation and context in which truths and lies are told" (Blair, Levine, & Shaw, 2010, p. 423). As we argue below, the same focus is apparent in current security screening practices.

Veracity testing offers an alternative approach to detecting deception that focuses, not on displayed behavioral characteristics of deceivers, but on the nature of the verbal exchange between the sender (the individual attempting to deceive) and the receiver (the individual attempting to detect deception). Recent laboratory studies have revealed six aspects of dyadic verbal exchanges that can discriminate deceivers from truth-tellers, which we outline below.

First, some of the most successful deception detection methods, which can yield up to 85% accuracy, use evidence-based techniques to determine veracity. For example, in interviews that use tactical and strategic information-gathering methods, the receiver first explores the sender's accounts, guided by information known to them before the interview, but without revealing to the sender what is known. The sender's responses are then compared with the known information, and challenged when inconsistencies arise by revealing the information. Tactical interviewing (TUE) uses a piece-by-piece approach to question, and then challenges accounts using known information items one at a time (see Dando & Bull, 2011; Dando, Bull, Ormerod, & Sandham, 2013). Strategic interviewing (SUE) typically involves revealing information and challenging receiver's accounts in bulk at the end of an interview once the receiver has answered all of the sender's questions (e.g., see Hartwig, Granhag, Stromwall, & Kronkvist, 2006; Vrij, Granhag, Mann, & Leal, 2009).

Second, questioning styles that elicit rich verbal accounts are also effective in discriminating between truth-tellers and liars (Milne & Bull, 1999; Oxburgh & Dando, 2011; Oxburgh, Myklebust, & Grant, 2010). Open questions do not constrain responses, but necessitate the provision of expansive answers. More important, answers to open questions commit passengers to an account of the truth concerning issues such as identity, background, and previous, current or future activities.

Third, tests of expected knowledge, which compare the content of what someone says with information already known, are useful for detecting deception (Blair, Levine, & Shaw, 2010). For example, if you claim to have studied at Oxford University, it would be reasonable to expect you to know how to travel on public transport from the train station to your college. Lack of knowledge and an inability to explain its absence, or a marked change in verbal behavior when providing answers, may suggest that the information supplied initially may not be veridical.

Fourth, interviewing methods that restrict the verbal maneuvering of deceivers are also shown to be effective (e.g., Dando & Bull, 2011; Taylor et al., 2013). Verbal maneuvering involves the strategic manipulation by deceivers of verbal content and delivery, which is intended to control a conversation to avoid detection. The quantity of verbalizations produced by deceptive individuals (measured in terms of number of words), and the information content of their verbalizations, tend to vary according to the nature of a verbal exchange. Specifically, deceptive individuals tend to be as verbose as truthful individuals when they are in control of the conversation (e.g., during early exchanges), and they tend to produce as much unsolicited information (and sometimes more) than truth-tellers. However, deceivers become less verbose and deliver less information than truth-tellers when their accounts are being challenged under questioning (Dando et al., 2013).

Fifth, procedures that raise the cognitive load faced by an interviewee typically yield better rates of discrimination between deceivers and truth-tellers (Walczyk, Igou, Dixon, & Tcholakian, 2013). For example, asking unanticipated questions during interviews has been shown to raise the cognitive load of deceivers more than truth-tellers, leading to higher detection rates (Vrij et al., 2009).

Sixth, speech content in response to questions can discriminate between truth-tellers and deceivers. For example, the cognitive interview (see Fisher & Geiselman, 1992), originally devised for interviewing witnesses and victims, has recently been modified for deception detection purposes and has been found to elicit noticeably different verbal responses from deceivers and truth-tellers (Morgan et al., 2013). Deceivers' response length, unique words, and type-token ratio (ratio of response length and token words) when answering prompts differ significantly to truth-tellers who speak longer, say more, and use more unique words than deceivers.

The six techniques described above provide building blocks for constructing an effective method for detecting deception during security interviews, and we describe their use in devising a composite approach to security screening below. Other interview approaches to detecting deception have been developed, such as the Reid technique (Inbau, Reid, & Buckley, 1986) and strategic questioning (Levine, Shaw, & Shulman, 2010). However, these are accusatory and confession-orientated, and as such are unsuitable for aviation security interviews, which are customer-focused and not suspect/persons of interest interviews.

Aviation Security Screening to Detect Deception

Most current aviation security procedures rely on the identification of behavioral indicators (e.g., British Security Industry Association, 2008; Reddick, 2004). A common method for screening airline passengers before embarking on long-haul flights involves the detection of "suspicious signs" during a short scripted interview between security agent and passenger (Martonosi & Barnett, 2006). In the interview the agent asks a series of security-related questions that are the same for every passenger. During questioning, agents look for indicators, which are typically behaviors associated with previous security incidents. These signs focus on aspects of a passenger's verbal and nonverbal behaviors, disposition (e.g., nervousness), and appearance (e.g., inappropriate dress for the intended trip) that may be indicators of deceit or threat.[1]

The psychological literature reveals potential problems with behavioral indicator approaches such as the suspicious signs method. In the context of aviation security screening, security agents will almost always be unfamiliar with the passengers that they screen, and so according to situational familiarity theory (Stiff et al., 1989), they are likely to resort to relying upon the kinds of nonverbal indicator that have been shown to be poor predictors of deception. Indeed, in the case of the suspicious signs method, agents are actively trained to do so.

In response to these concerns, we developed a new security screening method, which we call Controlled Cognitive Engagement (CCE). The name refers to the decision-making skills used by the security agent to control an interview so that a passenger provides information that can be tested for veracity. CCE embodies each of the six techniques shown in laboratory studies to improve deception detection rates: use of evidence; tests of expected knowledge; effective questioning styles; observation of verbal maneuvering; asymmetric cognitive loading; and changes in verbal behavior. Information revealed in the responses to open questions by the passenger in a CCE security interview is used by the agent to construct questions that provide tests of expected knowledge.

CCE is designed to raise the cognitive load faced by deceptive passengers (based on the assumption that passengers who pose a threat to a flight will have to be deceptive if questioned appropriately). At the same time, cognitive load is minimized for legitimate passengers, who experience what appears to be a friendly and informal conversation, albeit that the conversation is managed. That is, the security agent is seamlessly directing the course and progress of the conversation to fulfil specific psychological and practical objectives (e.g., Shepherd, 2007).

CCE is applied during a short interview between security agent and passenger. The interview procedure has three phases, involving a short phase of rapport-building followed by cycles of information-gathering and veracity-testing phases, but the questions within each phase are not scripted. Instead, agents use a process to derive in real time questions that cannot be predicted in advance by passengers. The process creates novel questions for each passenger, selecting combinations of topic (e.g., education, family, travel, work, skills, hobbies, and interests), temporality (past, current, and future), focus (self, others), information structure (indicative, spatial, relational, etc.), and question style (open for gathering a wide range of information, focused for gathering specific classes of information, and tests of specific knowledge). This process is trained as a question-design algorithm that is stepped through by the agent to create appropriate questions for each phase during the interview. An example interview is given in Table 1, showing the phases of the interview, and examples of question construction.

An initial phase of rapport building establishes a baseline for the passenger's verbal behavior, in which they respond to neutral questions that anyone, regardless of intent, could answer truthfully. In an information-gathering phase, passengers are then asked open questions created by the agent using the question design process described above. The account given by a passenger in response to an open question is then tested for veracity, using focused "test" questions that seek information the passenger should possess if their account is true. Agents assess the veracity of accounts from the answers given by passengers to questions. Although agents can access Internet resources using a hand held device to check the accuracy of some answers given by passengers, they are discouraged from doing so (and in the trial reported below, none did). Instead, they are trained to look for changes in the depth and quality of answers during the interview (i.e., as a deceptive passenger becomes aware that their account is being assessed or challenged). In particular, they look for inconsistencies within the interview, gaps in the passenger's expected knowledge, misalignments with the agent's prior knowledge, and changes in the passenger's behavior in response to test questioning (e.g., reduction in response length and information content). The coherence of the interview is maintained by bridging remarks made by the agent to make the dialogue seem natural. Furthermore, test questions may be preceded by one or more focused questions, typically beginning "who," "what," "where," "when," and "how," which open a topic detail that can then be addressed by a test question. Cycles of open and test questions repeat with different topics, temporal domains, foci, and information types, after which the agent makes a decision regarding the passenger's risk status.

Evaluating the Effectiveness of Aviation Security Screening Procedures

Recent research highlights the importance of context in assessing the effectiveness of methods for detecting mal-intent (Blair, Levine, & Shaw, 2010). Human behavior is inherently determined by the situation in which actions arise (Ross & Nisbett, 1991). To date, there have been no large-scale field trials of aviation security screening methods. Some proponents of behavioral indicators (e.g., Ekman, 2009) have noted the difficulty of trialing approaches in controlled studies that cannot involve genuine threat (Weinberger, 2010), because of the difficulty in a mock passenger study of creating the high stakes faced by perpetrators of real attacks. However, as Blair, Levine, and Shaw point out, the effectiveness of any method for detecting deception is likely to be influenced strongly by the context in which it is applied.

To address the gap in knowledge about how security screening methods fare in real task environments, we conducted the first in

[1] Ownership of, and access to, the suspicious signs method is restricted through national security legislation. Thus, the specific questions asked during a suspicious signs interview cannot be provided here. Further information regarding the method may be obtained from the authors.

Table 1
A Controlled Cognitive Engagement (CCE) Interview Protocol, Showing the Phases of the Interview, Example Features of Question Construction, and Example Questions That Might Be Asked by an Agent

Phase	Topic	Temporality	Focus	Information	Example question
Introduction					*Good morning, really sorry for the wait. How are you this morning?*
Rapport					*So how was your trip to the airport today, what with all this snow it must have been tough getting here?*
Bridging remark					*OK, as part of our security process, I'm just going to ask a few questions, to get to know you a little.*
Cycle 1					
Open	Education	Past-distant	Self	Location	*Can you tell me a little about your highest educational qualification? [Passenger answer includes - High School]*
Test					*Who was your high-school principal?*
Bridging remark					*Thanks. [looks at passport} I see you do a lot of travelling.*
Cycle 2					
Open	Family	Future-recent	Others	Plan	*Tell me a little about the family members you may be visiting in the next year [Passenger answer includes - Brother who lives in X]*
Focused					*How often do you visit him? [Passenger answer includes - Every year, at least once]*
Focused					*What part of X does he live in? [Pax gives locale Y]*
Test					*So, how long does it take to drive from the airport to Y?*
Bridging remark					*Ok that's not too long a trip, I guess.*
Cycle 3					
Open	Work	Current	Self	Function	*OK, so explain to me who you work for? [Pax answer includes - Company Z]*
Focused					*Ah, and what does Company Z do?*
Test					*Where is their main headquarters?*
Completion					*Thanks very much Sir/Madam, now you can proceed to the check-in desk.*

vivo evaluation of a suspicious signs method conducted in a real airport, and compared it with CCE. The study reported below provides a randomized-control, double-blind field trial of these aviation security-screening methods. We collected data at five international airports during routine security activities, in which mock and real passengers passed through a security interview before check-in for long-haul flights. Mock passengers were incentivized to attempt to pass through security undetected while giving untruthful answers during the screening interview.

We hypothesized that security agents using CCE would detect more mock passengers than agents using the suspicious signs method. We also hypothesized that interviews using the methods would be distinguished by the verbal behaviors of both passengers and agents: in CCE interviews, agents would speak less and passengers more, compared with suspicious signs interviews; CCE interviews would yield more information from passengers than suspicious signs interviews; and the methods would differ in the question types used by security agents. Finally, we hypothesized that deceptive passengers interviewed under the CCE method would show a reduction of verbal content (number of words and information items) in their answers as the interview proceeded.

Method

Participants

Security agents. Ninety-seven males and 65 females (M_{age} = 37.4, SD = 12.73) from a workforce of 866 staff participated as part of normal working but were free to withdraw from study participation at any time. Agents were aware that tests of screening effectiveness would be conducted during the 8-month trial, but were blind to the presence of mock passengers. Written consent was obtained from agents to audio-record interviews. All staff had received training in suspicious signs screening (2 weeks of classroom instruction and 1 week on-the-job training), and had an average of 4.9 years' experience (SD = 2.81) with the method.

Suspicious signs training has four main components: general security awareness raising, instruction in the kinds of signs (particularly nonverbal cues) that are believed to be indicators of deception, rote learning of a scripted sequence of questions concerning the passenger's travel itinerary and luggage, and procedures for conducting documentation checks (e.g., passport, ticket, visa; for an example program requirement see http://www.tsa.gov/stakeholders/commercial-airlines).

Staff were randomly allocated to suspicious signs (83 agents) and CCE (79 agents) methods. Staff allocated to suspicious signs received no further training. Staff allocated to CCE underwent CCE training, comprising 1 week of classroom instruction (lectures, role-play exercises, and continuous assessment), and 1 week on-the-job training. The classroom sessions comprised an introduction to myths and best known practices in deception detection, instruction on the rapport-building, information-gathering, and testing phases, an overview of decision-making and the evaluation of evidence, and a discussion of how to handle special cases (e.g., older, younger, and vulnerable passengers; groups of travelers; passengers with specific religious or cultural requirements). To pass the CCE training, agents had to complete the written classroom assessments with a pass mark of >80%. In addition, they were assessed during their final on-the-job training session, and were required to conduct three interviews in succession that conformed to best practice in CCE interviewing. All but two agents passed the CCE training.

Passengers. Two groups participated (113 male and 91 female in each group): mock passengers (mean age 45.7 years, $SD = 5.14$) and genuine passengers (mean age 43.5 years, $SD = 4.77$). Nationalities matched across groups were: American (43), British (116), German (16), Irish (5), Swiss (11), and Canadian (13). Mock passengers were recruited by advertisement. Each was paid £60 to participate, with an additional £60 if they passed through the security interview undetected as deceptive. During days when audio-recordings were taken, a researcher approached each passenger after their security interview, seeking permission for their interview to be included in the research. The sample of genuine passengers was selected from 7,000 security interviews audio-recorded during the evaluation, matched to mock passengers by age, gender, nationality, and first language.

Procedure

One week before participation, mock passengers provided full names, gender, date and place of birth, passport number, and home and work addresses. Each participant received an apparently valid ticket, an itinerary of flights and a deceptive cover story to be maintained during the security interview, and a ticketing record was created on the airline electronic database. Cover stories were unique and age- and gender-appropriate. For example, one participant (a male police officer from Hull) was told:

> You have worked as a telecoms engineer for the past three years in Northampton where you grew up. You are traveling to Chicago to take part in an international amateur fencing competition.

Participants were encouraged to undertake preparatory research to make their cover story convincing, by conducting Internet searches to gather information relevant to the cover story. On the second and fifth days of the week before participation, each participant took part in a telephone interview with a researcher, in which they were asked to give a free verbal account of their cover story and the information that they had accessed to enrich their relevant story knowledge. A threshold for successful telephone interview was set, requiring participants to maintain at least 2 min of uninterrupted verbal account of their cover story, and to supply at least four pieces of information that were not included in the original cover story. All participants met this threshold in both interviews. On the day they attended the airport, they were told to dress appropriately for their destination and travel story, and to bring appropriate luggage and a current passport.

Screening interviews took place before check-in for 88 long-haul flights. On arrival at the airport, each mock passenger met a researcher who used a written protocol to explain participation. No information was provided about security procedures. Mock passengers were directed individually (out of sight of security staff) to join genuine passengers queuing to pass through either the suspicious signs or CCE security process. They were then interviewed by a security agent. If an agent was not satisfied with the outcome of the interview (e.g., suspicious signs were detected but not resolved; the veracity of an account was not established), the passenger was identified for further screening under a separate mandatory process. Throughout the trial, security agents were fitted with digital audio recorders. Signs were posted informing passengers that security interviews may be recorded.

Results

Detecting Deception

Our principal interest was to determine which method was most effective for detecting deception (i.e., discriminating mock from genuine passengers). CCE agents identified 24 times more mock passengers (66%; 60% at Month 1 and 72% at Month 6) than suspicious signs agents (3%; 6% at Month 1 and 0% at Month 6), $\chi^2(1) = 13.091$, $p < .001$. Discrimination accuracy was assessed using a d' measure for each flight calculated from true positives (mock passengers) and false alarms (genuine passengers) identified as being deceptive. Agents using CCE, $M_{\mathrm{CCE}} = 3.016$, 95% CI [2.65; 3.38], showed higher discrimination accuracy than agents using suspicious signs, $M_{\mathrm{suspicious\ signs}} = 0.042$, 95% CI [−.34; .43], $F(1, 136) = 119.853$, $p = < .001$, $\eta_p^2 = .47$. The time of testing (Month 1 vs. Month 6) did not affect discrimination accuracy, all Fs $< .912$, all ps $> .404$.

Treating the genuine passengers who were identified as being deceptive as false alarms is a simplification, because it cannot be established whether they were mistakenly identified as deceptive or were actually being deceptive. However, base rates of identification of genuine travelers identified as being deceptive in the 6 months before the trial (1 in 1,247 passengers) did not differ during the trial with suspicious signs (1 in 1,219 passengers), or CCE (1 in 1,295 passengers), ps > 0.20.

Differences in Verbal Content

We hypothesized that the using CCE method would make it easier to detect deception because it generates more passenger verbalization than the suspicious signs method. To test this hypothesis, we analyzed the interview audio-recordings for the number of words spoken by each of the mock and genuine passengers, the number of words spoken by the agent, the number of questions asked by the agent, the types of questions asked (open, closed), and the temporal reference of questions (i.e., whether the question referred to an event in distant past, recent past, present, near future, or far future). In addition, we counted the number of unique information items (i.e., information unknown to the agent and unavailable through documentation) given in the verbal account of each of the mock and genuine passengers. Independent coders coded 20% selected from an initial sample of 202 transcripts. Intercoder reliability for number and types of questions, screener and passenger words, and information items was high: $r(82) = .827$, $p < .001$; $r(82) = .901$, $p < .001$; $r(82) = .988$, $p < .001$; $r(82) = .899$, $p < .001$; $r(82) = .841$, $p < .001$, respectively.

One possibility is that agents treated mock passengers differently from genuine passengers from the outset, irrespective of the interview method being used (i.e., agents recognized mock passengers as being different from genuine passengers, so changed their interview approach accordingly). As a manipulation check, between-subjects analysis of variances (ANOVAs) were conducted for each of our measures, including Passenger Type (mock vs. genuine) as a factor. These analyses revealed no significant main effects or interactions involving Passenger type for numbers of agent questions or words, all Fs $< .009$ and all ps $> .507$, passenger words or information items, Fs $< .179$, all ps $> .673$, or open, closed question types and temporalities, Fs $< .179$, all ps $> .308$.

It appears that, although agents using CCE were able to detect mock passengers whereas agents using suspicious signs could not, the way in which agents applied each interview method did not differ between mock and genuine passengers.

Examination of the frequencies of verbal behaviors (shown in Figure 1), averaged over passenger type, indicates differences between CCE and suspicious signs interviews. Passengers (both mock and genuine) screened using CCE uttered more words, CI 95% [254.54; 275.40], than those screened using suspicious signs, CI 95% [58.39; 79.25], $F(1, 400) = 683.543$, $p < .001$, $\eta^2 = .63$. Passengers screened using CCE also revealed more information items, $M_{\text{CCE items}} = 11.64$, CI 95% [11.23; 12.05], than passengers screened using suspicious signs, $M_{\text{Suspicious Signs items}} = 0.76$, CI 95% [0.36; 1.17], $F(1, 400) = 1379.924$, $p < .001$, $\eta^2 = .77$. CCE agents uttered fewer words, CI 95% [118.08; 131.51], than suspicious signs agents, CI 95% [316.20; 329.63], $F(1, 400) = 1686.806$, $p < .001$, $\eta^2 = .80$. Thus, the results confirm our hypothesis that CCE yields more verbal behaviors from passengers, regardless of whether they were mock or genuine.

Turning to the verbal behaviors of agents conducting the interviews, CCE agents asked fewer questions, CI 95% [10.35; 11.50], than suspicious signs agents, CI 95% [19.48; 20.64], $F(1, 400) = 486.089$, $p < .001$, $\eta^2 = .60$. CCE agents asked more open questions, CI 95% [3.06; 3.28], than suspicious signs agents, CI 95% [0.15; 0.37], $F(1, 400) = 1376.880$, $p < .001$, $\eta^2 = .76$, and their questions covered more temporal domains, CI 95% [2.10; 2.25], than those of suspicious signs questions, CI 95% [1.09; 1.24], $F(1, 400) = 355.066$, $p < .001$, $\eta^2 = .47$. However, CCE agents asked fewer closed questions, CI 95% [1.65; 2.04], $F(1, 400) = 1645.126$, $p < .001$, $\eta^2 = .80$, than suspicious signs agents, CI 95% [7.37; 7.76] (see Figure 2).

An important practical consideration is how long it takes to administer each interview method. To examine this issue, we measured the duration of each interview (in seconds, from agent introduction until directing the passenger toward check-in). As with other measures, no effects of Passenger Type (mock vs. genuine) were found for interview duration, all Fs < 3.278, all ps $> .071$. The difference in duration between CCE interviews ($M_{\text{CCE duration}} = 193.62$, $SD = 30.90$) and suspicious signs interviews ($M_{\text{suspicious signs duration}} = 186.95$, $SD = 35.30$), was not significant, $F = 2.064$, $p = .152$. Thus, regardless of whether passengers were mock or real, interview duration did not differ significantly according to which method was used.

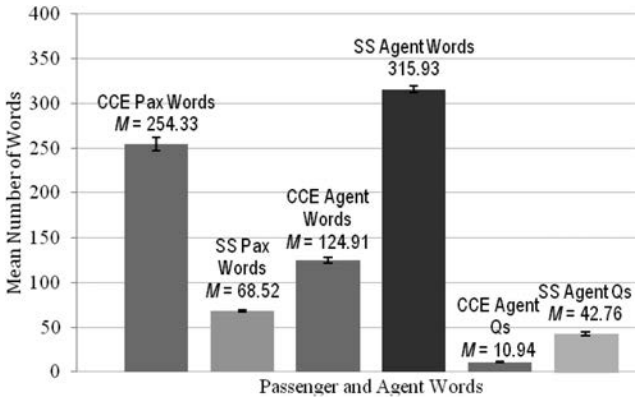

Figure 1. Mean number of words spoken by passengers (Pax—averaged over both mock and genuine) and agents (Controlled Cognitive Engagement: CCE and Suspicious Signs: SS) as a function of interview method ($N = 404$).

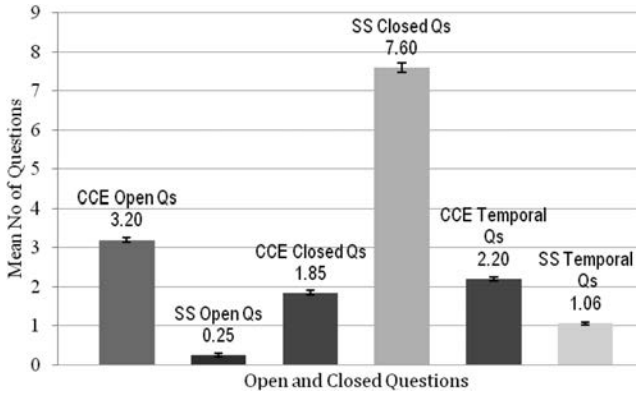

Figure 2. Mean question types asked by agents as a function of interview method ($N = 404$). CCE = Controlled Cognitive Engagement; SS = Suspicious Signs.

Change in Verbal Content

We also hypothesized that CCE is effective because it promotes tests of the veracity of passenger accounts, and these tests yield changes in the verbal behaviors of deceivers during an interview that can be used by security agents to discriminate them from truthful individuals. To test this hypothesis, we compared the number of words given by mock and genuine passengers in answer to the first and last open questions asked by a CCE security agent during an interview (i.e., comparing verbalizations before accounts have been challenged to verbalizations once deceptive individuals may have become aware that their accounts are being assessed). The number of words generated overall by mock passengers, $M_{\text{mock words}} = 24.80$, CI 95% [23.43; 26.27], and by genuine passengers, $M_{\text{genuine words}} = 26.54$, CI 95% [25.1; 27.9] did not differ significantly, $F(1, 199) = 2.96$, $p = .087$, $\eta^2 = .015$. For both mock and genuine passengers, the number of words uttered in answer to the first open question, $M_{\text{first words}} = 28.56$, CI 95% [27.10; 29.95] was significantly larger than the number uttered in answer to the last open question, $M_{\text{last words}} = 22.80$, CI 95% [21.85; 23.81], $F(1, 199) = 52.90$, $p < .001$, $\eta^2 = .21$. More important, the interaction between Passenger Type and Question was significant, $F(1, 199) = 74.341$, $p < .001$, $\eta^2 = .27$. Figure 3 illustrates the interaction, and shows that, while mock and genuine passengers uttered similar numbers of words in answers to the first open question of the interview, mock passengers uttered fewer words than genuine passengers in their answers to the last open question.

Similarly, we predicted that mock passengers would include fewer information items than genuine passengers in their answers to the focused questions used by agents to test the veracity of accounts. The number of information items generated overall by mock passengers, $M_{\text{mock items}} = 3.43$, CI 95% [3.19; 3.67], was lower than by genuine passengers, $M_{\text{genuine items}} = 4.30$, CI 95% [4.05; 4.54], $F(1, 199) = 24.7$, $p < .001$, $\eta^2 = .11$.

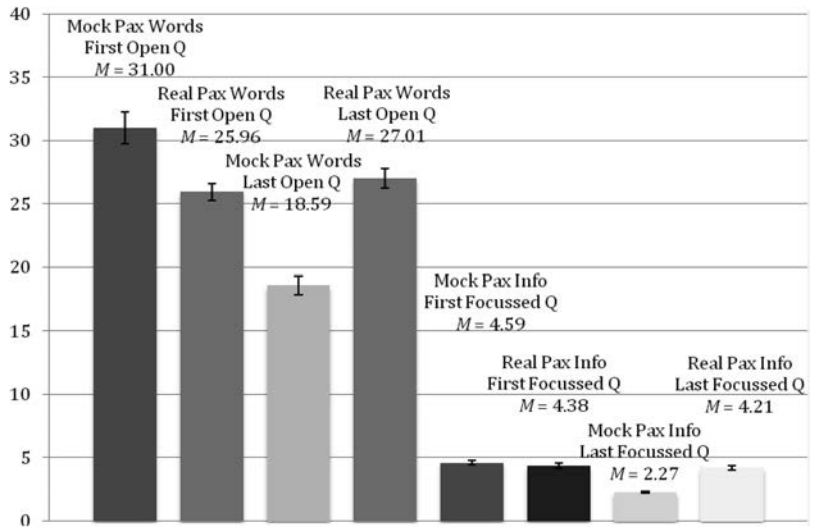

Figure 3. Mean number of words and information items as a function of passenger type (mock; genuine) and question position (first; last: $N = 404$).

For both mock and genuine passengers, the number of information items yielded in the first test cycle, $M_{\text{first items}} = 4.49$, CI 95% [4.24; 4.74] was significantly larger than the number yielded in the last test cycle, $M_{\text{last items}} = 3.24$, CI 95% [3.07; 3.41], $F(1, 199) = 87.056$, $p < .001$, $\eta^2 = .30$. Again, the interaction between Passenger Type and Question was significant, $F(1, 199) = 64.925$, $p < .001$, $\eta^2 = .25$. Figure 3 illustrates this interaction, and shows that, although mock and genuine passengers included similar numbers of information items in their answers to the first test cycle of the interview, mock passengers gave fewer items than genuine passengers in their answers to the last test cycle.

Discussion

The results of the field trial show a clear advantage for veracity testing over suspicious signs as a method for aviation security screening. CCE detected more mock passengers without increasing false alarm rates. With both mock and genuine passengers, CCE interviews yielded more passenger talk and information. At the same time, security agents produced less talk and asked fewer but more effective questions. The use of CCE changed the verbal behavior of deceptive passengers, whose answers became shorter and had less information content by the end of the interview, whereas the answers of genuine passengers did not change. By using an information-gathering approach, first asking open questions about unpredictable topics that vary in their temporal reference, followed by test questions that seek information an individual should possess if they are being truthful, it is likely that CCE minimized cognitive demand for legitimate passengers but increased it for deceivers (Beckmann, 2010). The failure of a suspicious signs approach to detect mock passengers is consistent with the poor performance of behavioral indicators found in laboratory studies of deception (see Bond & DePaulo, 2006; DePaulo et al., 2003), and extends this finding to a composite method (where more than one indicator is being sought) that is tested in a field setting.

The difference between suspicious signs and veracity testing approaches parallels a distinction in the decision-making literature between System 1 and System 2 modes of thinking (e.g., Evans, 2008; Kahneman, 2011). System 1 thinking uses cues in a task environment to trigger decision heuristics (e.g., Klein, 2004). System 2 thinking invokes more deliberative analytic decision-making and searches for counterexamples to initial inferences (e.g., Johnson-Laird, 2006). Security screening using the suspicious signs approach is analogous to System 1 thinking in using behavioral indicators to guide decision-making. Recognition of environmental cues and consequent retrieval of appropriate action sequences is a hallmark of expert decision-making, particularly in dynamic and time-critical domains (Klein, 2004; Schraagen, Militello, Ormerod, & Lipshitz, 2008). However, this kind of System 1 expertise develops from repeated exposure to cues in a relevant task environment. Security agents are rarely exposed to known incidents of deception, and cannot develop the kinds of automated expertise in cue recognition seen in other expertise domains. As a consequence, behavioral indicator methods for security screening necessarily comprise a rigid procedure in which the kinds of indicators to look for are prescribed and trained.

The scripted nature of a suspicious signs interview makes it difficult to use psychologically validated techniques such as tactical and strategic use of evidence methods (Dando & Bull, 2011; Dando et al., 2013; Hartwig et al., 2006), tests of expected knowledge (Blair et al., 2010), and unexpected questions (Vrij et al., 2009). Intuitive processing in deception detection of the kind promoted by a suspicious signs approach exacerbates truth and lie biases in deception judgments (Meissner & Kassin, 2002). Although the suspicious signs method is only one way of implementing behavior detection, we argue that any method that relies on behavioral indicators will yield disappointing results, because individual behavioral cues are weakly correlated with deception.

Veracity testing is more effective than suspicious signs screening because it encourages a System 2 mode of thinking, in which the consistency of an individual's account is assessed analytically. Veracity testing directly addresses the problem identified in Porter, Gustaw, and ten Brinke's (2010) Dangerous Decisions Theory. CCE leads security agents to assess the content of individual's account, rather than relying on intuitive judgments about passenger appearance and disposition. Every passenger is different, and CCE enables agents to adapt their interviewing to reflect these differences.

Although the difference between the methods in terms of effectiveness is striking, some limitations of the study must be borne in mind. For example, we used financial incentives to motivate participants to try and evade detection. It is moot as to whether it is ultimately possible to simulate deception in high stakes environments. However, on the few occasions when mock passengers were detected by agents using the suspicious signs approach, they were typically identified on the basis of nervousness, suggesting that the manipulation was successful to some degree. Furthermore, like all field studies, the current study necessarily raises concerns of internal validity, in terms of other potential mediating factors (e.g., passenger volume, test time, etc.). We ameliorated these concerns in four main ways: we undertook a comparison of performance at two different times during the study under different passenger load volumes; we reduced investigator effects by not being present and by the agents being blind to the nature and timing of mock passenger participation; by standardizing instructions across conditions; and equating demand characteristics by equalizing financial incentives across conditions. Nonetheless, a large part of the value of the study lies in it being conducted in a real security setting.

It is possible that CCE might become less effective over time. For example, individuals with mal-intent might try to "break" the method through extensive practice (e.g., dry runs, rehearsal with unanticipated questioning, and careful preparation of scripted knowledge). A subset of 20 of the current participants went through the CCE method on more than one occasion (in some cases, up to four times over the course of the trial—only data from their first run are reported above). The detection rates were 16/20 for this subset at Time 1, and 14/20 at Time 2 ($p > .05$). Of the 8 participants who went through the method four times, 6 were detected on the first trial, 4 on the second trial, 6 on third trial, and 5 on the fourth trial. The sample of repeat screenings is small. However, the fact that individuals who were intent on maximizing their financial gain by evading detection were unable to do so, despite repeated prior exposure to the nature of the procedure, goes some way to assuring the longevity of the CCE method.

There are additional benefits of a veracity testing approach. Behavioral indicators associated with previous terrorist events may not predict future events; CCE identifies deceit in real time, allowing discovery of new kinds of threat. The unpredictability of CCE questioning breaks any lie script a deceiver might prepare (Von Hippel & Trivers, 2011), which reduces opportunities for reverse engineering of a security-screening method and subsequent evasion by perpetrators (Chakrabarti & Strauss, 2002). In contrast, the suspicious signs method comprises a fixed sequence of closed questions to which responses can be rehearsed (e.g., during "dry runs"). Finally, passive observation of passenger behaviors carries a risk of selective profiling that may disadvantage some ethnic, gender, and age groups. CCE is applied equally to all passengers, avoiding inappropriate profile-based biases.

Our results have implications for practitioners, both in security screening, and more generally for professional lie catchers such as police officers and court officials. In line with recent psychological research on deception detection (e.g., Reinhard et al., 2011), security methods should aim to maximize the amount of verbal content received from the sender (the potential deceiver), and should direct the receiver (e.g., the security agent) to focus on verbal content rather than nonverbal cues. We suggest that deception detection methods that rely on observation are unlikely to be as effective as those that include a dyadic verbal exchange. In contrast to current practice, we propose that security agents should not be trained to identify specific behaviors associated with deception, because this kind of training is likely to amplify the negative effects of situational familiarity (Stiff et al., 1989). Instead, agents should rely upon preexisting knowledge of social interaction to identify when a passenger's behavior changes during an interview.

The low failure rate among security agents taking the CCE training course suggests that veracity testing methods lie within the competence of most adults: few of the agents had more than a high-school level education, yet they were able to implement the method successfully. Equipped with the right interview methods, humans can be entrusted with the task of security screening. We suggest caution in the use of technologies to replace or support interview processes. In our pilot work to design the method, we equipped agents with PDAs for conducting Internet searches to check information given by passengers. However, agents were reluctant to use the facility, and when they did, the process of conducting external checks interrupted the interview, which could offer a deceptive passenger thinking time in which to prepare answers to any challenge of their account. In our view, relying on veracity testing and noting the behavior change that arises under challenge is sufficient for making reliable judgments about potential security threats.

In closing, we note that the sensitivity of CCE for detecting deception is unique. Most studies of deception detection use base rates of around 50:50 deceivers to truth-tellers. Here, high rates of deception detection were obtained with a base rate of less than 1:1000 mock to genuine passengers. Thus, a more positive picture emerges of the contribution that can be made by psychological research to the protection of public safety than previously thought.

References

Beckmann, J. F. (2010). Taming a beast of burden: On some issues with the conceptualization and operationalization of cognitive load. *Learning and Instruction, 20,* 250–264. http://dx.doi.org/10.1016/j.learninstruc.2009.02.024

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research, 36,* 423–442. http://dx.doi.org/10.1111/j.1468-2958.2010.01382.x

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10,* 214–234. http://dx.doi.org/10.1207/s15327957pspr1003_2

British Security Industry Association. (2008). *Detecting behaviour to prevent aviation attacks*. Retrieved from http://www.bsia.co.uk/aboutbsia/news/newsarticle/N4BCQB63655

Chakrabarti, S., & Strauss, A. (2002). Carnival booth: An algorithm for defeating the computer-based passenger screening system. *First Monday, 10*. Retrieved from http://firstmonday.org/issues/issue7_10/chakrabarti/index.html

Dando, C. J., & Bull, R. (2011). Maximising opportunities to detect verbal deception: Training police officers to interview tactically. *Journal of Investigative Psychology and Offender Profiling, 8,* 189–202. http://dx.doi.org/10.1002/jip.145

Dando, C. J., Bull, R., Ormerod, T. C., & Sandham, A. (2013). Helping to sort the liars from the truth-tellers: The gradual revelation of information during investigative interviews. [Advance online publication]. *Legal and Criminological Psychology, n/a*. http://dx.doi.org/10.1111/lcrp.12016

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129,* 74–118. http://dx.doi.org/10.1037/0033-2909.129.1.74

Ekman, P. (2009). Lie-catching and micro-expressions. In C. Martin (Ed.), *The philosophy of deception. Oxford: Oxford University Press*. http://dx.doi.org/10.1093/acprof:oso/9780195327939.003.0008

Evans, J. S. B. T. (2008). Dual-processes accounts of reasoning, judgment, and social congnition. *Annual Review of Psychology, 59,* 255–278. http://dx.doi.org/10.1146/annurev.psych.59.103006.093629

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Charles C Thomas.

Hartwig, M., Granhag, P. A., 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. http://dx.doi.org/10.1007/s10979-006-9053-9

Inbau, F. E., Reid, J. E., & Buckley, J. P. (1986). *Criminal interrogations and confessions*. Baltimore, MD: Williams and Walkins.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin.

Klein, G. (2004). *The power of intuition: How to use your gut feelings to make better decisions at work*. New York, NY: Currency.

Levine, T. R. (2010). A few transparent liars: Explaining 54% accuracy in deception detection experiments. [Sage.]. *Communication Yearbook, 34*, 40–61.

Levine, T. R., Kim, R. K., & Blair, J. P. (2010). (In)accuracy at detecting true and false confessions and denials: An initial test of a projected motive model of veracity judgments. *Human Communication Research, 36,* 82–101. http://dx.doi.org/10.1111/j.1468-2958.2009.01369.x

Levine, T. R., Shaw, A., & Shulman, H. C. (2010). Increasing deception detection accuracy with strategic questioning. *Human Communication Research, 36,* 216–231. http://dx.doi.org/10.1111/j.1468-2958.2010.01374.x

Martonosi, S. E., & Barnett, A. (2006). How effective is security screening of airline passengers? *Interfaces, 36,* 545–552. http://dx.doi.org/10.1287/inte.1060.0231

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. http://dx.doi.org/10.1023/A:1020278620751

Milne, R., & Bull, R. (1999). *Investigative interviewing: Psychology and practice*. West Sussex: Wiley.

Morgan, C. A., Rabinowitz, R. G., Hilts, D., Weller, C. E., & Coric, V. (2013). Efficacy of modified cognitive interviewing, compared to human judgments in detecting deception related to bio-threat activities. *Journal of Strategic Security, 6,* 100–119. http://dx.doi.org/10.5038/1944-0472.6.3.9

Oxburgh, G., & Dando, C. J. (2011). Interviewing witnesses and suspects: Where now in our search for the Truth? *The British Journal of Forensic Practice, 13,* 135–147.

Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech, Language & the Law, 17,* 45–66.

Porter, S., Gustaw, C., & ten Brinke, L. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law, 16,* 477–491. http://dx.doi.org/10.1080/10683160902926141

Reddick, S. R. (2004). Point: The case for profiling. *International Social Science Review, 79,* 154–156.

Reinhard, M. A., Sporer, S. L., Scharmach, M., & Marksteiner, T. (2011). Listening, not watching: Situational familiarity and the ability to detect deception. *Journal of Personality and Social Psychology, 101,* 467–484. http://dx.doi.org/10.1037/a0023726

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York, NY: McGraw-Hill.

Schraagen, J. M. C., Militello, L., Ormerod, T. C., & Lipshitz, R. (Eds.). (2008). *Macrocognition and naturalistic decision making*. Aldershot, United Kingdom: Ashgate Publishing Limited.

Shepherd, E. (2007). *Investigative interviewing: The conversation management approach*. Oxford: Oxford University Press.

Stiff, J. B., Miller, G. R., Sleight, C., Mongeau, P., Garlick, R., & Rogan, R. (1989). Explanations for visual cue primacy in judgments of honesty and deceit. *Journal of Personality and Social Psychology, 56,* 555–564. http://dx.doi.org/10.1037/0022-3514.56.4.555

Taylor, P. J., Dando, C. J., Ormerod, T. C., Ball, L. J., Jenkins, M. C., Sandham, A., & Menacere, T. (2013). Detecting insider threats through language change. *Law and Human Behavior, 37,* 267–275. http://dx.doi.org/10.1037/lhb0000032

United States Government Accountability Office. (2011). *Aviation security: TSA is taking steps to validate the science underlying its passenger behavior detection program, but efforts may not be comprehensive*. GAO-11-146T.

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34,* 1–16. http://dx.doi.org/10.1017/S0140525X10001354

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior, 33,* 159–166. http://dx.doi.org/10.1007/s10979-008-9143-y

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches. [Advance online publication]. *Frontiers in Psychology, 4,* 14. http://dx.doi.org/10.3389/fpsyg.2013.00014

Weinberger, S. (2010). Airport security: Intent to deceive? *Nature, 465,* 412–415. http://dx.doi.org/10.1038/465412a

Received June 14, 2014
Revision received August 19, 2014
Accepted September 17, 2014 ■