APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1375

Successful Detection of Verbal and Visual Concealed Knowledge Using an RT-Based Paradigm

TRAVIS L. SEYMOUR* and JESS R. KERLIN

University of California, Santa Cruz, USA

SUMMARY

An increasing number of researchers are exploring variations of the Concealed Knowledge Test (CKT) as alternatives to traditional 'lie-detector' tests. For example, the response times (RT)-based CKT has been previously shown to accurately detect participants who possess privileged knowledge. Although several studies have reported successful RT-based tests, they have focused on verbal stimuli despite the prevalence of photographic evidence in forensic investigations. Related studies comparing pictures and phrases have yielded inconsistent results. The present work compared an RT-CKT using verbal phrases as stimuli to one using pictures of faces. This led to equally accurate and efficient tests using either stimulus type. Results also suggest that previous inconsistent findings may be attributable to study procedures that led to better memory for verbal than visual items. When memory for verbal phrases and pictures were equated, we found nearly identical detection accuracies.

Although the polygraph-based Control Questions Test (CQT, also called the 'lie-detector' test) (Reid, 1947; Reid & Inbau, 1977) has had considerable applied success and has been shown to be accurate in several studies (for recent reviews, see Honts, Raskin, & Kircher, 2002; Kleiner, 2002), others have questioned its theoretical basis (e.g. Elaad, 2003; Steinbrook, 1992), test accuracy (e.g. Ben-Shakhar, Bar-Hillel, & Lieblich, 1986; Cross & Saxe, 1992; Iacono & Lykken, 2002; Kleinmuntz & Szucko, 1982; Lykken, 1998) and resistance to countermeasures (e.g. Honts, Raskin, & Kircher, 1987; Iacono & Lykken, 1997; Kleiner, 2002; Rosenfeld, Soskins, Bosh, & Ryan, 2004). Recently, the National Research Council (2003) completed a comprehensive review pointing out shortcomings of the CQT as well as other available measures of deception. This report not only concludes that current approaches need further testing and refinement, but that new alternative tools should be pursued (National Research Council, 2003).

TESTING 'GUILTY' AND 'CONCEALED' KNOWLEDGE

To address some of the shortcomings of 'lie-detection', Lykken (1959) introduced an alternative approach based on the detection of 'guilty' knowledge. The standard Guilty Knowledge Test (GKT) uses a paradigm where participants are presented with a question

*Correspondence to: Travis L. Seymour, Department of Psychology, University of California, 374 Social Sciences 2, Santa Cruz, CA 95064, USA. E-mail: nogard@ucsc.edu

and a series of answer choices, one of which refers to the actual crime under investigation. For example, 'the criminal in this case left an article of clothing at the crime scene. Was it (a) a blue hat, (b) a black glove, (c) a white shirt, (d) a brown shoe or (e) a green tie?' Similar to 'lie-detection' paradigms, a polygraph machine is often used to measure participants' physiological responses (i.e. heart rate, respiratory rate and electrodermal response) to test questions. A consistently stronger physiological response to the crime-related items than to control items indicates knowledge of crime details. Laboratory studies using the GKT report showing higher overall classification accuracy than typical lie-detection paradigms (e.g. Bashore & Rapp, 1993; Ben-Shakhar & Elaad, 2003; MacLaren, 2001), although some reports dispute this interpretation (cf. Honts et al., 2002). Although using the polygraph to index recognition is generally successful, physiological measures in the GKT can be influenced by individual differences in responsivity (Lykken, 1998), or simple physical countermeasures (Ben-Shakhar & Dolev, 1996).

Because he did not have a 'way to observe recognition directly' (Lykken, 1998, p. 285), Lykken proposed the use of a polygraph machine in conjunction with the GKT. Later, Farwell and Donchin (1991) proposed a more direct measure of recognition using electroencephalograms measured at the scalp. This approach involves tracking changes in brain electrical potential that follow stimulus presentation called 'event-related potentials' (ERP) and represents a refinement of previously developed techniques (Donchin & Coles, 1988; Fabiani, Gratton, Karis, & Donchin, 1987; Rosenfeld, Angell, Johnson, & Qian, 1991; Rosenfeld & Bessinger, 1990; Rosenfeld, Cantwell, Nasman, Wojdac, Ivanov, & Mazzeri, 1988). A substantial positive ERP deflection approximately 300 milliseconds after the presentation of a familiar stimulus appears to serve as a distinctive indicator of a participant's recognition of relevant or salient information (e.g. Bashore & Rapp, 1993; Fabiani et al., 1987). In the Farwell and Donchin (1991) paradigm, which we refer to as the Concealed Knowledge Test (CKT)[1] to distinguish it from the aforementioned GKT, participants learned a set of two-word phrases (e.g. 'Blue Coat') and then used this information to commit a 'mock crime'. After a delay, participants memorised a new set of similar phrases (e.g. 'Green Tie'), called *Targets,* and were subsequently given a speeded recognition task. They were asked to acknowledge their familiarity of the recently learned *Target* phrases by responding 'Old', and their lack of familiarity with novel *Irrelevant* phrases by responding 'New'. A third category of *Probe* phrases was also presented. During 'innocent' blocks, Probe phrases were new and thus similar to Irrelevants, while Probes presented during 'guilty' blocks were taken from the mock crime. During 'guilty' blocks, participants were asked to conceal their familiarity with Probes by quickly and accurately responding 'New' to these items as if they were novel Irrelevants. By examining whether ERPs on Probe trials were more closely correlated with those on Target trials (i.e. showed a substantial P300) or Irrelevant trials (i.e. lacked a substantial P300), Farwell and Donchin (1991) were able to reliably determine when participants' responses were produced during a 'guilty' or 'innocent' block. This ERP-based paradigm was tested in laboratory and applied contexts and produced an overall classification accuracy[2] of 87.5%,

[1]This test was originally called the GKT (Lykken, 1959) or technique (Lykken, 1960), but has also been named the concealed information test (Verschuere, Crombez, & Koster, 2004), and the concealed knowledge test (e.g. Honts, Devitt, Winbush, & Kircher, 1996). We use the term concealed knowledge test because of the paradigm's use in non-forensic settings (e.g. Jacoby, 1991), and because, as Honts et al. (1996) point out, 'knowledge cannot be guilty, and knowledge, not the actual information, is what is being concealed'.

[2]Following Lykken (1998), accuracy is calculated by summing the hit rate (correct detections in those with concealed knowledge) and the correct rejection rate (correct detections in those without concealed knowledge) and dividing this by two.

DOI: 10.1002/acp

similar to the average success reported for the polygraph-based CKT (Bashore & Rapp, 1993; Ben-Shakhar & Elaad, 2003; MacLaren, 2001).

Seymour, Seifert, Mosmann, and Shafto (2000) examined the possibility of using only response times (RT) and accuracy to detect concealed knowledge. In this study, participants memorised several details required to send an incriminating e-mail message (e.g. an e-mail address, an operation codename, a file name and a street location). They were asked to e-mail another student (e.g. 'Dale Spence') the location (e.g. 'Perch Street') of a file (e.g. 'Rain File') containing instructions for hacking the university database to permit the alteration of one's grades. Seymour et al. (2000) used a CKT paradigm similar to that reported by Farwell and Donchin (1991), but added a 1000 millisecond response deadline and a detection algorithm that compared Probe and Irrelevant RT distributions as well as differences in error rates. This test led to a 96.5% classification accuracy. Follow-up experiments revealed that, even when motivated to do so, participants could not equate their RT on Probe and Irrelevant trials during familiar-Probe blocks. Because this classification scheme is simpler and more accurate than the ERP procedure, the RT-CKT appears to be a highly successful paradigm for detecting concealed knowledge.

Success using RT measures appears limited to this particular CKT paradigm. For example, Verschuere, Crombez, De Clercq, & Koster (2004) reported a modified dot-Probe paradigm in which familiar Probe and unfamiliar Irrelevant pictures were briefly flashed and replaced by a dot stimulus. Participants ignored the flashed pictures and manually responded to the dot stimulus (e.g. left button for '..' and right button for ':'). Although processing the flashed pictures was irrelevant to performing the dot classification, the presence of familiar Probe pictures captured participants' attention causing them to classify the dots more slowly than on neutral trials where both pictures were Irrelevant. However, in a follow-up study using this paradigm, RT differences did not emerge (Verschuere et al., 2004).

Another paradigm that yields inconsistent RT results is the Emotional Stroop task (EStroop; Williams, Mathews, & MacLeod, 1996). In this task, participants were presented with emotional and neutral words and asked to name their ink colour under time pressure. Compared to neutral stimuli (e.g. 'boot'), responses were typically slower when the underlying word was personally relevant, threatening or highly emotional (e.g. 'murder'). Gronau, Ben-Shakhar, and Cohen (2005) reported two RT-based EStroop tests: One in which the critical words were taken from a previously committed mock crime, and another that used personally relevant items such as the participant's name. When naming the ink colour of personally relevant stimuli, participants were slower compared to neutral words. However, mock crime-related words did not produce such a difference. Engelhard, Merckelbach, and van den Hout (2003) also failed to find RT differences in a similar EStroop-based detection task.

VERBAL VS. VISUAL STIMULI IN THE CKT

Although successful RT-based tests have been reported using the CKT paradigm (Allen, Iacono, & Danielson, 1992; Farwell & Donchin, 1991; Rosenfeld et al., 2004; Seymour et al., 2000), they have largely ignored non-verbal stimuli despite the prevalence of photographic evidence in forensic cases. However, visual stimuli have been used with various GKT paradigms. For example, in a pupil-dilation based GKT using a combination of names and photographs as stimuli, Lubow and Fein (1996) found that visual and verbal

DOI: 10.1002/acp

stimuli contributed equally to detection. However, in two similar studies using polygraph-based GKTs, differences in electrodermal response depended on whether Probe stimuli were verbal or visual. In a blocked design, participants were asked to study either schematic faces (simple line drawings) or verbal descriptions (e.g. detailed profiles including personality, occupation, hobby and hometown) that varied along several dimensions. Participants were later tested using either the previously studied stimulus or a slightly altered version. Verbal descriptions lead to a greater mean electrodermal response than schematic faces (Ben-Shakhar & Gati, 1987; Gati & Ben-Shakhar, 1990). Thus, previous work does not provide a clear prediction for the relative success of verbal and visual stimuli in the GKT. Furthermore, no previous work has reported the success of visual stimuli using the CKT paradigm. To address this, we examined the detection efficiency of an RT-based CKT using pictures as stimuli and compared its performance to one using verbal phrases as stimuli. It is not clear why the Ben-Shakhar studies show an enhanced response for verbal stimuli compared to visual stimuli, while Lubow and Fein (1996) showed no effect of modality. However, one possibility is that personal history descriptions led to richer memory representations than schematic faces, which in turn caused differences in participants' later recognition of those items (Craik & Lockhart, 1972). Thus, we predicted that if level of encoding was kept constant, CKTs using visual and verbal stimuli would produce similar detection accuracies. Alternatively, if encoding is equated and the verbal CKT outperforms a visual test, as in the Ben-Shakhar and Gati studies, then we would also conclude that concealed knowledge detection is more successful using verbal stimuli.

To test this hypothesis, we administered the RT-based CKT in two stimulus modality conditions: one using pictures of human faces as stimuli and the other using two-word verbal phrases. Because it was critical that the visual and verbal study procedures lead to comparable levels of encoding for each stimulus type, we also collected data from a post-test task designed to compare Picture and Phrase memory.

METHOD

Participants

Participants were 64 undergraduate students enrolled in an introductory psychology course at the University of California and were randomly assigned to a Picture or Phrase condition. Students participated in the experiment for course credit.

Materials

Equipment

Experimental stimuli were displayed on a 17″ display (640×480 at 85 Hz) and participants were seated 0.6 m away. A Cedrus RB-420 button box was used to collect manual responses and relay them to a computer which recorded RT and accuracy.

Stimuli

Stimuli in the Picture condition consisted of 126 greyscale digital photographs of faces (half male), 7.0 cm × 9.5 cm in size, from the Aberdeen Psychological Image Collection (Hancock, 2004). This set included happy, angry and neutral versions of each of 42 unique

DOI: 10.1002/acp

faces. However, emotional pictures were used only during Probe study; faces presented for Target study and all test pictures were of neutral expression. For each participant, 36 unique neutral faces were randomly selected and, for 6 of these faces, the corresponding happy and angry variations were also selected. Stimuli in the Phrase condition consisted of two-word phrases from a set of 72 phrases (e.g. 'Perch Street') developed by Farwell and Donchin (1991).

Procedure

Each participant performed a concealed knowledge task twice during the experimental session; once in the familiar-Probe block, where the Probe items were those studied earlier in the session, and once in the unfamiliar-Probe block, where probes were taken from the Probe set studied by a different participant. Block order was counterbalanced by participant so that for half of the participants, the familiar-Probe block occurred first. Participants were randomly assigned to either a face picture or verbal phrase stimulus condition. Thus, the experimental factors were Probe Familiarity (familiar-Probe vs. unfamiliar-Probe blocks), manipulated within-participant and Stimulus Modality (Pictures vs. Phrases), manipulated between participants. For both stimulus conditions, the procedure consisted of the following sequence of tasks: a Probe study task, a distractor task, two test blocks and memory post-tests for Target and Probe stimuli. Each test block consisted of a Target study task, and a stimulus classification task.

Probe study task

In the Picture condition, participants saw 36 faces selected at random from the stimulus set described above. Six of these faces were chosen at random to be *Probe* faces. Each Probe face was presented for 10 seconds, and participants were asked to study its overall shape and features. On each study trial, the study picture featured either a neutral, happy (smiling) or angry (frowning) expression. After the face-study period, the screen was cleared and participants were asked questions about the size, shape and appearance of various facial attributes (e.g. 'How big was the person's nose? Small, medium or large', or 'What type of haircut did the person have? Crew, mullet, etc.'). Because they could not predict which question would be asked for any particular face (questions were chosen randomly with replacement), participants were encouraged to study each face carefully. Following the attribute question, the same person's face would be presented with either the same emotional expression as shown during the face-study period, or with one of the two other possible emotional expressions. Participants were asked to press a button marked 'Same' if this picture depicted the same emotional expression as the study picture; otherwise the 'Different' button was pressed. For this task, accuracy was stressed over speed. Same vs. different trials were randomised, along with the particular emotional valence of the image shown on 'different' trials. This first phase of Probe study (study, attribute question and same/different emotion judgement) was repeated three times and counterbalanced so that after three iterations of this phase, participants had been exposed to all three versions of each Probe face. The order in which faces were studied was re-randomised before each repetition of this phase.

Following the first phase of Probe study, participants completed a face rating task. During this task, all six Probes were presented in three randomly ordered sequences. On each trial, both emotional versions of each face were presented horizontally adjacent to one another on either side of the screen's midpoint. For each display, participants were asked to

DOI: 10.1002/acp

rate the apparent attractiveness, honesty or age of each person. Attractiveness, honesty and age ratings were blocked, and counterbalanced for each participant.

In the Phrase condition, 36 two-word phrases were chosen for each participant as described above, and 6 were randomly chosen to be Probe phrases. Participants were shown all six phrases in a randomly ordered list and asked to commit them to memory in order to carry out an e-mail-based 'mock crime' scenario (Seymour et al., 2000). This information was used to login to the university computer system and send another student an electronic message containing details of how to hack into the university registrar grade database. In order to achieve this, participants first memorised the specific information they would use to commit the crime. Training software presented six critical items that participants were to memorise. For example, participants may be asked to send a message to *Phil Jenks* regarding *Operation Cow* indicating that they would find someone wearing a *White Shirt* on *Perch Street* who would hand them the *Rain File*. Following one presentation of all the information, participants completed a cued-recall task; for example, 'street name' was presented as a prompt for 'Perch Street'. The participants typed in an answer at each prompt. This study-test sequence was repeated three times.

Next, participants were instructed to execute the instructions they had just memorised and actually commit the computer 'crime'. A computer display was presented with what appeared to be an interface to a university e-mail client. Each participant, following the studied scenario, logged into the e-mail system and sent the prepared message. Though the task appeared realistic to participants, the e-mail message was stored to disk and not actually sent. The task ended once participants successfully 'sent' the electronic message and logged out.

Distractor task

Following the Probe study task, participants completed a 10-minute survey designed to occupy working memory and prevent rehearsal of the Probe items. The task consisted of 11 challenging mathematical word problems (taken from Patalano & Seifert, 1994).

Target study

After the distractor task, participants learned a new set of six *Target* faces or phrases by using procedures similar to the one described for the Probe study task. Unlike during Probe study, participants in the Picture condition did not rate attractiveness, honesty or age of Target faces, and participants in the Phrase condition were not asked to use the Targets in any mock-crime scenario. Thus, the face ratings task used during Probe study and absent during Target study offers a slightly more elaborated memory for Probes which participants could later use to distinguish one list from another (Craik & Lockhart, 1972). Similarly, in an applied use of this paradigm, Probes taken from the crime scene will likely be more elaborated in the guilty suspect's mind than Target items chosen during the examination. This is also similar to the relationship between relevant and control items in the polygraph-based CQT (Lykken, 1998).

Classification task

The classification task was repeated in two trial blocks, each consisting of a Target-study phase in which participants studied six new Targets, and a speeded recognition task featuring a random sequence of Target, Probe and Irrelevant phrases. The 6 Probes (previously studied in the familiar-Probe block, and novel in the unfamiliar-Probe block) and 6 familiar Targets were combined with 24 new Irrelevant items and presented in a

DOI: 10.1002/acp

random order. This same set of 36 items was re-randomised and presented a total of 3 times during the block for a total of 108 trials (18 Targets, 18 Probes and 72 Irrelevants). In both the familiar-Probe and unfamiliar-Probe blocks, the same instructions were given. Participants were instructed to acknowledge their familiarity of Targets by pressing a button labelled 'Old', and to accurately reject new Irrelevant items by pressing a button labelled 'New'. In addition, participants were asked to conceal their knowledge of familiar Probe items by responding 'New' to these items, just as quickly and accurately as on Irrelevant trials. Thus, 'New' was the correct response for both Irrelevant and Probe items, and 'Old' was the correct response for Target items only. In the unfamiliar-Probe block, participants would not be able to distinguish Probes and unfamiliar Irrelevants, thus responding 'New' to these stimuli accurately represents their lack of recognition. Before each stimulus was displayed, the word 'Ready' was displayed for 500 milliseconds, followed by a fixation cross for 500 milliseconds. Each stimulus remained on the screen until a response was made.

Before and throughout the classification task, participants were urged to respond as quickly and accurately as possible. On trials in which a participant's response spanned more than 1500 milliseconds, the message 'Too Slow' was displayed for 1000 milliseconds before the next stimulus was displayed; otherwise, no feedback was given between trials. The inter-trial interval was randomly varied between 800 and 1200 milliseconds to prevent response preparation and rhythmic response patterns. After each 36 trials (i.e. 3 times per block), participants were shown a feedback screen indicating separate accuracy rates for Target items (referred to as 'items on the list you just studied') and Probe/Irrelevant items (referred to as 'items not on the list you just studied'). They were also shown the number of 'Too Slow' responses, and when it was greater than zero, they were reminded to respond both quickly and accurately. On each trial, the RT and accuracy of each response was recorded.

Memory post-test

Our goal for the Probe study procedure in both the Picture and Phrase conditions was to replicate the general structure previously reported in Seymour et al. (2000) and involved studying the six Probes, completing a cued-recall test of these items, and finally completing a task designed to help elaborate that information in participants' memory. The Phrase condition afforded the closest replication and also used the 'mock crime' used in this previous study. A cued-recall study task was designed to ensure that participants were indeed learning the Probe phrases, however, this recall format was not feasible in the Picture condition. Instead, participants were asked to demonstrate their memory of each face picture indirectly by recalling an aspect of some random facial feature, and identifying whether a newly presented version of the face was the same or different emotion as the one they studied. Finally, in an effort to match the degree of elaboration provided by the e-mail scenario in the Phrase condition, we used a set of rating tasks in the Picture condition. Despite this attempt to equate Phrase and Picture study, it is possible that the resulting Probe memory in the two conditions were significantly different which would confound the interpretation of any differences between tests.

To address this, we conducted unspeeded post-tests of both Probes and Targets following the classification task. Participants in the Phrase condition were asked separately to recall as many Probe phrases ('first list') and then Target phrases ('second list') as possible. The order of these questions was counterbalanced. Participants in the Picture condition could not be asked to recall images, and were thus given a randomised 6×3 picture grid showing

DOI: 10.1002/acp

the six Probe, six Target and six randomly selected Irrelevant faces. They were separately asked to indicate which faces were from the Probe list and the Target list. No feedback was given about each selection and the question order was counterbalanced.

RESULTS

Five participants were excluded from the analysis for failing to follow instructions. Debriefing of these participants revealed a misunderstanding of instructions such that the 'Old' button was used for any familiar stimulus.

Phrase CKT results

Mean RT data for the phrase CKT is shown in Figure 1. Assessing the concealed knowledge effect in this paradigm involves averaging across participants (as opposed to averaging across all trials) and comparing Probe and Irrelevant distributions. Prior to participant averaging, individual responses faster than 300 milliseconds and slower than 2000 milliseconds (<1%) were excluded as outliers and not used in the analyses. A 2 (Stimulus Type: Probe vs. Irrelevant) × 2 (Probe Familiarity: familiar-Probe block vs. unfamiliar-Probe block) repeated-measures ANOVA was performed on RT and revealed a main effect of Stimulus Type, $F(1,24) = 43.01$, $p < .001$, MSE = 5663, $\eta_p^2 = .64$, a main effect of Probe Familiarity, $F(1,24) = 27.26$, $p < .001$, MSE = 14074, $\eta_p^2 = .53$ and Stimulus Type × Probe Familiarity interaction, $F(1,24) = 41.48$, $p < .001$, 5581, $\eta_p^2 = .63$. A similar set of analyses could not be performed on the accuracy data due to ceiling effects associated with rejecting unfamiliar items (see Table 1). However, using a *Z*-test for

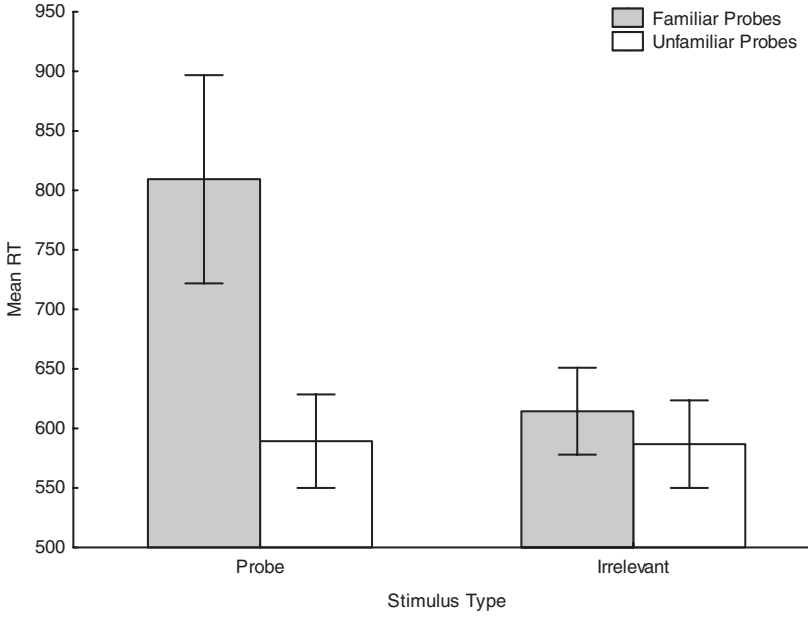

Figure 1. Mean correct RT in milliseconds by Probe Familiarity and Stimulus Type in the Phrase condition. Bars represent standard deviations

DOI: 10.1002/acp

Table 1. Mean per cent correct by Probe Familiarity condition and Stimulus Type for each stimulus modality

	Stimulus Type			
Block	Probe		Irrelevant	
Phrases				
Familiar-Probe (%)	76	(30)	98	(6)
Unfamiliar-Probe (%)	99	(3)	99	(2)
Pictures				
Familiar-Probe (%)	73	(25)	99	(2)
Unfamiliar-Probe (%)	97	(5)	97	(4)

Note: Values enclosed in parentheses represent standard deviations.

proportions we confirmed the main predictions that Probe accuracy (76%) would be lower than Irrelevant accuracy (98%) in the familiar-Probe condition, $Z = -3.19$, $p < .001$, but equal (99%) in the unfamiliar-Probe condition, $Z < 1$, $p = .98$, was confirmed. Thus, a concealed knowledge effect is revealed where, during the unfamiliar-Probe block, participants were unable to distinguish Probes from Irrelevants. However, during the familiar-Probe block, responses to Probes were considerably slower and less accurate on average. This pattern is similar to previous reports (Farwell & Donchin, 1991; Rosenfeld et al., 2004; Seymour et al., 2000).

Picture CKT results

Mean RT data for the picture CKT is shown in Figure 2. A 2 (Stimulus Type) × 2 (Probe Familiarity) repeated-measures ANOVA performed on RT revealed a main effect of Stimulus Type, $F(1,33) = 72.50$, $p < .001$, MSE = 6070, $\eta_p^2 = .69$, a main effect of Probe Familiarity, $F(1,33) = 32.66$, $p < .001$, MSE = 9916, $\eta_p^2 = .50$ and a Stimulus Type × Probe Familiarity interaction, $F(1,33) = 51.87$, $p < .001$, MSE = 5454, $\eta_p^2 = .61$. As with phrases, participants rarely made errors when rejecting unfamiliar faces (see Table 1). However, Probe accuracy (73%) was lower than Irrelevant accuracy (99%) in the familiar-Probe condition, $Z = -4.07$, $p < .001$, but equal (97%) in the unfamiliar-Probe condition, $Z < 1$, $p = .96$. Thus, similar to the Phrase condition, the typical concealed knowledge effect is revealed in the Picture condition, where participants were slower and made more errors when rejecting familiar Probes. Surprisingly, both the RT and accuracy means for Probes and Irrelevants in the Picture condition are within 5% of the data found in the Phrase condition (compare Figures 1 and 2). We tested this observation using a $2 \times 2 \times 2$ mixed-model ANOVA, in which Modality (Phrase vs. Pictures) served as the between-subjects variable, and Stimulus Type (Probe, Irrelevant) and Familiarity (familiar-Probes vs. unfamiliar-Probes) served as within-subjects variables. This analysis revealed no significant main effects or interactions involving modality (all Fs < 1). This may suggest that similar processing took place in both the Picture and Phrase paradigms.

Comparison of detection accuracy

Overall, mean RT and accuracy for the Phrase- and Picture-based tests looked similar. Both the overall magnitude and the interaction between Stimulus Type and Probe familiarity

DOI: 10.1002/acp

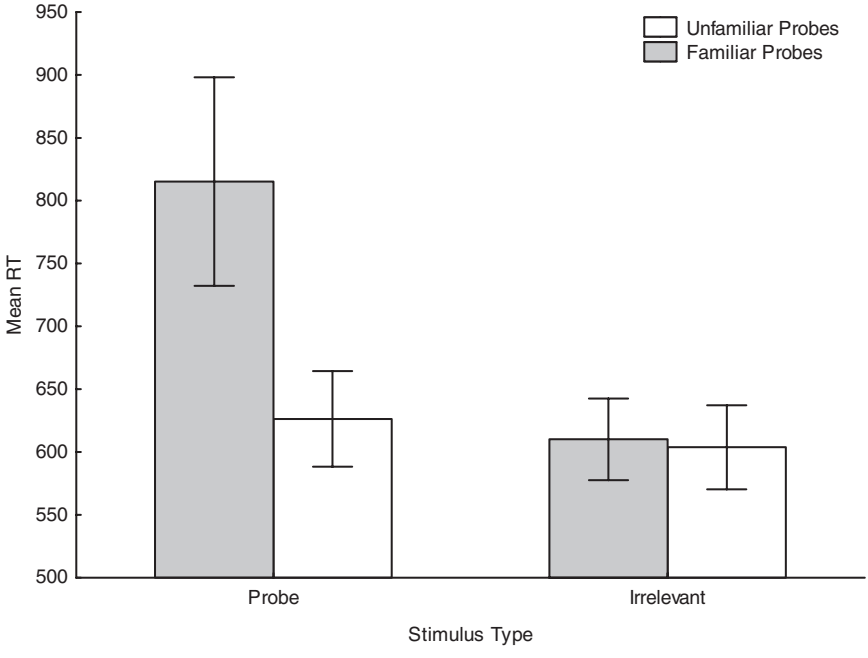

Figure 2. Mean correct RT in milliseconds by Probe Familiarity and Stimulus Type in the Picture condition. Bars represent standard deviations

were nearly identical. Because of this, we expected that detection accuracy for each measure would be comparable as well. Because it examines exactly how consistent the effects are for each participant, this type of analysis is more informative than ANOVA. To determine whether a participant's responses were taken from a familiar-Probe block or an unfamiliar-Probe block, we compared their Probe and Irrelevant responses. If Probes and Irrelevants appear to come from different underlying distributions, knowledge of the Probes is indicated. Probe and Irrelevant distributions that do not differ significantly indicate participants' ignorance of the Probe items. Although the data reported here show strong concealed knowledge effects for both RT and accuracy, it is possible that the effect for an individual participant will show up only in their RT *or* accuracy, limiting the effectiveness of either measure in isolation. Although in most cases slower RT *and* increased errors are expected, Seymour et al. (2000) reported an individual participant detection procedure designed to use only a participant's own data and examines the three main ways in which Probe and Irrelevant distributions are likely to differ. If these distributions differ significantly (Bonferroni corrected alpha: Shaffer, 1995) on RT variance (F-test for variances), number of errors (Fisher's Exact 2×2 Test), or the shape of RT cumulative probability distributions (Kolmolgorov–Smirnov Test: Kotz, Johnson, & Read, 1983), then the response is assumed to have come from the familiar-Probe block. When this occurs for Probe responses collected during the familiar-Probe block, a hit (i.e. correct detection) is registered. A significant difference on any one of these three statistical tests for Probe responses during the unfamiliar-Probe block indicates a false alarm (incorrect detection). Thus, for each participant we used this procedure to analyse data from both their familiar-Probe block (to determine hits and misses) and their

DOI: 10.1002/acp

unfamiliar-Probe block (to determine false alarms and correct rejections). In the Phrase condition, this analysis yielded a hit rate of .88 (22 hits and 3 misses) and a false-alarm rate of .04 (1 false alarm out of 25). For the Picture condition, the hit rate was .91 (31 hits and 3 miss) and the false-alarm rate was .03 (1 false alarm out of 34). A *Z*-test for proportions suggests that these detection rates did not differ from one another, $Z < 1$, and were similar to the .93 hit rate and .01 false-alarm rate reported by Seymour et al. (2000). Overall test accuracy was 94% for phrases and 95% for pictures.

A more complete indication of the fitness of a test can be produced by completing a receiver operating characteristic (ROC) analysis (Green & Swets, 1966).[3] Similar to our individual participant analysis procedure we entered both RT and accuracy (although not variance) into the ROC analysis. To do this, we created a composite CKT score by adding each participant's RT effect (Probe RT minus Irrelevant RT) to 100 times their accuracy effect (Irrelevant accuracy minus Probe accuracy). Note that *a* is a measure of the area underneath an ROC curve and typically ranges from .5 (poor efficiency) to 1 (perfect efficiency). In the present experiment, the areas underneath the ROC curves for the Phrase and Picture CKTs were 0.99 ($a > .5$: $Z = 72.69$, $p < .001$) and 0.95 ($a > .5$: $Z = 14.13$, $p < .001$) respectively. The detection efficiency for the Phrase- and Picture-based tests did not statistically differ, $Z = 1.41$, $p = .16$.

Post-test results

The post-test results allowed us to determine whether memory for pictures and phrases was comparable, or whether one modality led to better recollection than another. During the post-test, participants in the Phrase condition were asked to recall which items had been studied on the Probe and Target lists separately. Mean per cent correct source memory for Probes, 84% (SD = 18%), was not statistically different from Targets, 83% (SD = 11%), $t(24) = .07$, $p = .94$. We used a similar post-test to evaluate memory in the Picture condition. Correct source memory for Probe faces, 91% (SD = 12%) was greater than for Target faces, 80% (SD = 14%), $t(33) = 3.73$, $p < .01$. Despite this statistical difference, we note that both performance levels represent strong list discrimination. On average, participants misattributed approximately .6 out of 6 Probe items and 1.2 out of 6 Target items. The critical post-test comparison is between conditions. Our goal was to design Probe and Target study procedures that led to similar memory elaborations. Comparing post-test memory performance between conditions revealed no difference for Probes, $t(54) = -1.69$, $p = .09$, or Targets, $t(54) = 1.00$, $p = .32$. This suggests that participants' overall memory of pictures and phrases was comparable and may have influenced the similar detection accuracies we found for the visual and verbal versions of the CKT.

[3]The ROC curve considers the criterion or score that will be used to separate 'familiar' and 'unfamiliar' decisions. The resulting hit rate of each cutoff score is plotted against the false-alarm rate that would accompany that cutoff. To distinguish a 'good' test from a poor one, the area under this function (denoted as *a*) is considered and typically ranges from .5 to 1. With a poor test, the function follows the diagonal and each change in the cutoff score increases the hit rate and the false-alarm rate equally, resulting in an area of .5. With an ideal test, varying the cutoff value to increase the hit rate does not increase the false-alarm rate. If increasing the cutoff score progressively increases the hit rate, but does not affect the false-alarm rate, the underlying area would be 1. An area of an individual test's area can be evaluated for its distance from the .5 boundary, and the area under the ROC curve for two or more tests can be compared. If two tests yield similar ROC curves, they would be expected to yield roughly similar proportions of hits and false alarms (or 'detection efficiency').

DOI: 10.1002/acp

GENERAL DISCUSSION

Consistent with previous reports (Allen et al., 1992; Farwell & Donchin, 1991; Rosenfeld et al., 2004; Seymour et al., 2000), we found that RT and accuracy can be used to efficiently detect concealed knowledge. Furthermore, equally accurate detection was observed using either pictures or phrases as stimuli. This was consistent with our prediction, but in contrast to previous studies reporting detection differences for verbal and visual stimuli (Ben-Shakhar & Gati, 1987; Gati & Ben-Shakhar, 1990). In light of the present post-test results showing no significant difference in participants' memory as a function of stimulus modality, it is likely that participants in studies reporting enhanced detection for words may have had better memory for those items. Likewise, the lack of modality differences reported by Lubow and Fein (1996) may have been due to a similar level of encoding for verbal and visual stimuli. Thus, the present results suggest that pictures can be successfully used as stimuli in the RT-based CKT, and that if participants are sufficiently familiar with the critical test items, a Picture-based test can be just as accurate as one using verbal stimuli. Due to the prevalence of photographic evidence in modern forensics, accurate detection using both verbal and visual stimuli is an important demonstration for this paradigm.

The similar detection accuracy for phrases and pictures may appear surprising because only the Phrase condition used a mock crime. This suggests that feelings of 'guilt' or anxiety are not required to show the concealed knowledge effect. This is consistent with nearly identical non-applied paradigms showing similar results with benign stimuli. In this work, the effect is attributed to failure in list source memory (Jacoby, 1991) or to response competition (Seymour, 2001). According to these accounts, emotionality may enhance the differential response to Probe items, but is not strictly required. Similarly, Seymour and Fraynt (submitted for publication) have shown that in the RT-CKT paradigm, an elaborative Probe study procedure led to more stable detection over time (up to 1 week) than a more shallow study procedure. Thus, depth of encoding (Craik & Lockhart, 1972) may be more important than emotionality *per se*.

One limitation of the present study is the lack of direct comparisons with SCR-based tests using similar stimuli as well as the consideration of how countermeasures may attenuate the effect. Because using EDR measures requires long inter-trial intervals, and the use of fewer trials to avoid habituation effects, we did not simultaneously collect RT and EDR in the current study to maintain compatibility with previous reports. However, in ongoing research (Seymour & Verschuere, submitted for publication) designed to facilitate the optimal collection of both RT and EDR in a single CKT, our data suggest that the RT performs at least as well as EDR and is more resistant to countermeasures and habituation than previously suggested (Gronau et al., 2005; Rosenfeld et al., 2004).

Applied implications of the present results may also be limited by possible divergence between our participants and those suspected of committing actual crimes. This criticism applies to all laboratory research of mechanical detection tests. Although laboratory studies are a critical complement to field work, it has been suggested that laboratory rewards for 'beating' the test lack the real world consequences associated with tests given in the field. Specifically, some researchers have argued that laboratory tests may overestimate the size of the concealed knowledge effect examiners can expect in the field (Ben-Shakhar & Elaad, 2003; Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Elaad, 1990; Gronau et al., 2005). Although this conclusion is based on consistently larger effect sizes in laboratory settings than in the field, examining the influence of this

DOI: 10.1002/acp

difference on detection efficiency is less common. However, a recent study by Pollina, Dollins, Senter, Krapohl, and Ryan (2004) replicated this effect size difference, but showed that it did not influence test accuracy or diagnosticity. Similarly, Ben-Shakhar and Elaad (2003) conducted a large meta-analysis of GKT studies and found a significant difference in effect size when 'highly motivated' participants ($d = 1.76$) were compared to those with 'low motivation' ($d = 1.34$), but not on their respective test efficiencies ($a = .82$ and .80, respectively). We nonetheless believe that an essential next step for the RT-CKT is verification in field settings and are encouraged by at least one successful field report using the ERP-based CKT (Farwell & Donchin, 1991). Two successful court cases also reported using the ERP-based test (Farwell, 1999, 2000), however, peculiarities of how the test was applied in these cases may severely limit their implications (Rosenfeld, 2005).

The present data also have implications for the use of concealed knowledge paradigms more generally. While our results suggest that the RT-CKT can successfully detect concealed knowledge, we found identical effect sizes using procedures with and without mock crimes. This suggests that it may not be able to distinguish between those who are guilty and aware of crime details from those who are innocent but aware of this information. For example, consider that successful applied use of any CKT requires that some critical information be identified for use as Probes. In forensic cases, a subset of crime scene details not publicly released may be used. If it can be guaranteed that these items are known to only investigators and the perpetrator, then a test using this information may indeed determine whether the examinee is the guilty party. We note that it may not be necessary to conceal a large number of items from the public for use in the CKT. Although a test using 5 to 6 Probe items is ideal and yields an average test efficiency of .89, in many cases it can very be difficult to identify more than 3 (Podlesny, 2003). However, the Ben-Shakhar and Elaad (2003) GKT meta-analysis suggests that average efficiency for tests using between 1 and 4 items is still .82. Indeed, across 80 studies using between 1 and 25 Probes, the Pearson correlation between number of critical items and effect size was only .35, and the relationship between number of Probes and test efficiency was only .27.

In non-forensic cases such as national or corporate espionage investigations, Probes may be taken from classified material and presented to those unauthorised suspects believed to have viewed this information. While a CKT constructed from these materials can successfully indicate whether the unauthorised examinee has viewed privileged information, it may not reveal whether malicious intent was involved. While no existing test can accomplish this, previous research using polygraph measures have shown some success in distinguishing those who are 'guilty' from those who are 'innocent but aware' of crime details when familiarity with actions, rather than knowledge is stressed (Ben-Shakhar, Gronau, & Elaad, 1999; Bradley, MacLaren, & Carle, 1996; Bradley & Warfield, 1984). Although a promising approach, in some cases researchers report a significantly high false-alarm rate for innocent participants who are aware of crime information (e.g. Bradley et al., 1996). Despite this, such an approach may ultimately increase the applied contexts in which all variants of the CKT can be used. Without a way to differentiate the perpetrator from those who are innocent but aware of crime details, the CKT may be most useful for identifying or eliminating suspects than in pinpointing the guilty party.

We have shown that the RT-based CKT can be equally successful using both phrases and pictures as stimuli. The use of CKTs as evidence in legal and forensic settings has strong theoretical and empirical support (Ben-Shakhar, Bar-Hillel, & Kremnitzer, 2002) and variations of the CKT have already meet the Daubert criteria of legal admissibility in two recent court cases (Farwell, 1999, 2000; Makeig, 2002).

DOI: 10.1002/acp

ACKNOWLEDGEMENTS

The authors thank Aimee Fitzgerald, Becky Fraynt, Anna Kurtz, Lauren Ogren and all those who kindly agreed to participate in this study. We would also like to thank Colleen Seifert, Chris Baker and Josh Gaunt for their comments on earlier drafts.

REFERENCES

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, *29*, 504–522.

Bashore, T. R., & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin*, *113*, 3–22.

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: Reconsidering the use of the guilty knowledge technique in court. *Law and Human Behavior*, *26*, 527–541.

Ben-Shakhar, G., Bar-Hillel, M., & Lieblich, I. (1986). Trial by polygraph: Scientific and juridical issues in lie detection. *Behavioral Sciences & the Law: Special Issue: Psychology in Law Enforcement*, *4*, 459–479.

Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effect of mental countermeasures. *Journal of Applied Psychology*, *81*, 273–281.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*, 131–151.

Ben-Shakhar, G., & Gati, I. (1987). Common and distinctive features of verbal and pictorial stimuli as determinants of psychophysiological responsivity. *Journal of Experimental Psychology: General*, *116*, 91–105.

Ben-Shakhar, G., Gronau, N., & Elaad, E. (1999). *Leakage of relevant information to innocent examinees in the GKT: An attempt to reduce false-positive outcomes by introducing target stimuli*. US: American Psychological Assn.

Bradley, M. T., MacLaren, V. V., & Carle, S. B. (1996). Deception and nondeception in guilty knowledge and guilty actions polygraph tests. *Journal of Applied Psychology*, *81*, 153–160.

Bradley, M. T., & Warfield, J. F. (1984). Innocence, information, and the Guilty Knowledge Test in the detection of deception. *Psychophysiology*, *21*, 683–689.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the Guilty Knowledge Test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied*, *9*, 261–269.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*, 671–684.

Cross, T. P., & Saxe, L. (1992). A critique of the validity of polygraph testing in child sexual abuse cases. *Journal of Child Sexual Abuse*, *1*, 19–33.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral & Brain Sciences*, *11*, 357–427.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology*, *75*, 521–529.

Elaad, E. (2003). Is the inference rule of the 'control question polygraph technique' plausible? *Psychology, Crime & Law*, *9*, 37–47.

Engelhard, I. M., Merckelbach, H., & van den Hout, M. A. (2003). The Guilty Knowledge Test and the modified Stroop task in detection of deception: An exploratory study. *Psychological Reports*, *92*, 683–691.

Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 2, pp. 1–78). Greenwich, CT: JAI Press.

DOI: 10.1002/acp

Farwell, L. A. (1999). Brain fingerprinting test of J. G. Retrieved 6 June 2004, from http://www.brainwavescience.com/GrinderForensicReport.php

Farwell, L. A. (2000). Brain fingerprinting test on T. H. RE: State of Iowa vs. T. H. Retrieved 14 June 2004, from http://www.brainwavescience.com/HarringtonForensicReport.php

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ('lie detection') with event-related brain potentials. *Psychophysiology*, *28*, 531–547.

Gati, I., & Ben-Shakhar, G. (1990). Novelty and significance in orientation and habituation: A feature-matching approach. *Journal of Experimental Psychology: General*, *119*, 251–263.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, N.Y.: John Wiley.

Gronau, N., Ben-Shakhar, G., & Cohen, A. (2005). Behavioral and physiological measures in the detection of concealed information. *Journal of Applied Psychology*, *90*, 147–158.

Hancock, P. (2004). Psychological image collection at stirling. Retrieved 1 September 2000, from http://pics.psych.stir.ac.uk/

Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology*, *33*, 84–92.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology*, *1*, 241–247.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2). St. Paul, MN: West Publishing.

Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology*, *82*, 426–433.

Iacono, W. G., & Lykken, D. T. (2002). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2, pp. 483–538). St. Paul, MN: West Publishing.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, *30*, 513–541.

Kleiner, M. (Ed.). (2002). *Handbook of polygraph testing*. San Diego, CA: Academic Press.

Kleinmuntz, B., & Szucko, J. J. (1982). On the fallibility of lie detection. *Law & Society Review*, *17*, 85–104.

Kotz, S., Johnson, N. L., & Read, C. B. (1983). Kolmogorov-Smirnov statistics. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 393–396). New York, NY: John Wiley & Sons.

Lubow, R. E., & Fein, O. (1996). Pupillary size in response to a visual Guilty Knowledge Test: New technique for the detection of deception. *Journal of Experimental Psychology: Applied*, *2*, 164–177.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*, 385–388.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*, 258–262.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. New York, NY: Plenum Press.

MacLaren, V. V. (2001). A qualitative review of the Guilty Knowledge Test. *Journal of Applied Psychology*, *86*, 674–683.

Makeig, T. H. (2002). Brief of Amicus Curiae: Dr. Lawrence A. Farwell in support of appellant T. H. Retrieved 6 June 2004, from http://www.brainwavescience.com/Amicus%20Brief.php

National Research Council. (2003). *The polygraph and lie detection*. Washington, D.C.: National Research Council.

Patalano, A. L., & Seifert, C. M. (1994). Memory for impasses during problem solving. *Memory & Cognition, 22*, 234–242.

Podlesny, J. A. (2003). A paucity of operable case facts restricts applicability of the Guilty Knowledge Technique in FBI criminal polygraph examinations. *Forensic Science Communications*, *5*, 20–37.

DOI: 10.1002/acp

Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J., & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of Applied Psychology*, *89*, 1099–1105.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law, Criminology and Police Science*, *37*, 542–547.

Reid, J. E., & Inbau, F. E. (1977). *Truth and deception* (2nd ed.). Baltimore: Williams & Wilkins.

Rosenfeld, J. P. (2005). 'Brain Fingerprinting': A critical analysis. *The Scientific Review of Mental Health Practice*, *4*. Available on: http://www.fbi.gov/hq/lab/fsc/backissu/july2003/podlesny.htm.

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J.-H. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, *28*, 319–335.

Rosenfeld, J. P., & Bessinger, G. T. (1990). Feedback-evoked P300 responses in lie detection. *Psychophysiology*, *27*, S60.

Rosenfeld, J. P., Cantwell, B., Nasman, V. T., Wojdac, V., Ivanov, S., & Mazzeri, L. (1988). A modified, event-related potential-based Guilty Knowledge Test. *International Journal of Neuroscience*, *24*, 157–161.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*, 205–219.

Seymour, T. L. (2001). A EPIC model of the 'guilty knowledge effect': Strategic and automatic processes in recognition. *Dissertation Abstracts International: Section B: The Sciences & Engineering*, *61*, 5591.

Seymour, T. L., & Fraynt, B. R. (submitted for publication). Time lag and question choice in the concealed knowledge test. *Unpublished manuscript*.

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess 'guilty knowledge'. *Journal of Applied Psychology*, *85*, 30–37.

Seymour, T. L., & Verschuere, B. (submitted for publication). Detection of reaction-times versus skin conductance in the detection of concealed information. *Unpublished manuscript*.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Steinbrook, R. (1992). The polygraph test: A flawed diagnostic method. *New England Journal of Medicine*, *327*, 122–123.

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. W. (2004). Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology*, *41*, 461–466.

Verschuere, B., Crombez, G., & Koster, E. H. W. (2004). Orienting to guilty knowledge. *Cognition & Emotion*, *18*, 265–279.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, *120*, 3–24.

DOI: 10.1002/acp