Journal of Applied Psychology
2000, Vol. 85, No. 1, 30-37

0021-9010/00/$5.00 DOI: 10.1037//0021-9010.85.1.30

Using Response Time Measures to Assess "Guilty Knowledge"

Travis L. Seymour and Colleen M. Seifert
University of Michigan

Michael G. Shafto
Cognitive Science Associates

Andrea L. Mosmann
University of Michigan

How can a suspect's guilt or innocence be reliably tested? The validity of the polygraph, which measures changes in physiological arousal during a "guilty knowledge" test, is controversial (e.g., T. R. Bashore & P. E. Rapp, 1993; T. P. Cross & L. Saxe, 1992; D. T. Lykken, 1998; J. P. Rosenfeld, 1995; R. Steinbrook, 1992). One alternative to the polygraph examines event-related potentials recorded during a memory interference task (L. A. Farwell & E. Donchin, 1991). The present study extended this paradigm to determine whether response times (RTs) can accurately identify participants possessing specific guilty knowledge. Results from Experiment 1 showed that RT alone can reliably discriminate "guilty" from "innocent" participants. Experiments 2a and 2b indicated that an RT-based paradigm is more resistant to strategic manipulation than previously suggested (Farwell & Donchin, 1991). This RT-based paradigm may be a viable alternative to the polygraph for detecting guilty knowledge.

How can suspects be tested to reliably ascertain their guilt or innocence? A common method of determining whether someone is concealing information is through the use of a polygraph-based "lie-detector" test. The polygraph uses changes in physiological measures of arousal (e.g., breathing rate, blood pressure, and skin conductance) as an index to the emotional impact of a participant's responses. A suspect's responses to crime-related information are presumed to be untruthful when they are correlated with higher levels of arousal than evident in control questions. Despite numerous studies questioning its effectiveness (e.g., Bashore & Rapp, 1993; Cross & Saxe, 1992; Furedy & Heslegrave, 1988; Rosenfeld, 1995; Steinbrook, 1992), polygraph evidence is admissible in many U.S. states (when all parties consent).[1] In addition, the polygraph is still widely used in the areas of domestic disputes, workplace management, drug testing, and law enforcement investigations.

The preferred method for presenting questions during polygraph examination is the Guilty Knowledge Technique (GKT; Lykken, 1959, 1960, 1998), in which suspects are presented with questions related to a crime in an attempt to uncover any privileged information they may possess. For example, the question "What color was the getaway car?" would be followed by a series of response choices including the actual color of the car in question. Presumably, to an innocent participant, all color choices would be equally arousing; however, knowledge of the car's true color may produce differential physiological responses to the correct color than to other colors. If consistent differential responses occur for crime-related alternatives, the suspect is considered "guilty." However, the validity and reliability of the polygraph as a dependent measure continue to be controversial (Ben-Shakhar, Bar-Hillel, & Lieblich, 1986; Bradley & Warfield, 1984; Furedy, 1991; Furedy & Heslegrave, 1988; Kleinmuntz & Szucko, 1982).

As an alternative, event-related brain potentials (ERPs) have been proposed as a measure resistant to covert manipulation by suspects (Rosenfeld & Bessinger, 1990; Rosenfeld et al., 1988). ERPs are measured via scalp electrodes to detect changes in electrical patterns across the cerebral cortex related to the presentation of a stimulus. In particular, a pattern of electrical activity related to the recognition of a familiar but infrequent stimulus can be reliably detected. This "oddball" paradigm (Fabiani, Gratton, Karis, & Donchin, 1987) involves asking participants to classify a series of stimuli into two categories, one of which occurs less frequently than the other. A distinctive positive electrical potential occurs approximately 300 ms after an item from the low-frequency category appears as a stimulus (Donchin & Coles, 1988). This electrical potential, called the P300, appears to reflect surprise or interest and is affected by both the frequency of items in the less frequent category and their relevance to the task (Fabiani et al., 1987).

In a series of experiments, Farwell and Donchin (1991) used this oddball paradigm to test the value of ERPs in detecting "guilty

Travis L. Seymour, Colleen M. Seifert, and Andrea L. Mosmann, Department of Psychology, University of Michigan; Michael G. Shafto, Cognitive Science Associates, Ann Arbor, Michigan.

We thank Bill Gehring, Gail McKoon, David Meyer, Andrea Patalano, Roger Ratcliff, Brian Ross, and J. E. Keith Smith for assistance and helpful suggestions. We also thank all those who kindly agreed to participate in the study.

Correspondence concerning this article should be addressed to Travis L. Seymour, Department of Psychology, 525 East University, University of Michigan, Ann Arbor, Michigan 48109-1109. Electronic mail may be sent to nogard@umich.edu.

[1] Recent judgments by the Supreme Court (*United States v. Scheffer*, 1998) may further limit the use of the polygraph as trial evidence.

knowledge." Their participants "committed" one of two mock "crimes" after studying associated information (e.g., where to go, whom to meet, and what information to exchange). In a subsequent experimental session, participants memorized a new list of phrases ("targets"). They were then tested with a series of phrases and asked to classify each as either one of the previously studied target phrases or a new ("irrelevant") phrase. Because the targets occurred much less frequently than the irrelevant phrases, the authors predicted that a P300 would occur on trials in which a target (old) phrase was presented but not on trials with an irrelevant (new) phrase.

Farwell and Donchin (1991) also added a small number of "probe" phrases consisting of information from the mock crime that participants had committed. Participants were not informed about this third category and were expected to correctly reject its members as irrelevant (i.e., not from the target list). However, the researchers predicted that participants would recognize probe items as familiar on the basis of their crime knowledge and would therefore produce P300s, just as with stimuli in the target category. Participants were assessed on two separate test blocks: one in which participants were innocent (had no knowledge of the probe items) and one in which the probe items were taken from the participants' mock crime.

Farwell and Donchin (1991) found that probe trials containing information from the mock crime produced reliable P300s, but probe trials containing information from a mock crime participants did not commit failed to produce P300s (as did irrelevant trials). Comparison of ERPs during probe trials with ERPs during irrelevant trials allowed discrimination of blocks in which participants had guilty knowledge. From these results, it appears that detection of P300 responses in this oddball paradigm may be a viable alternative to the polygraph.

Farwell and Donchin (1991) collected another potential measure of participants' responses to test items: the time necessary to classify each test item as a target or irrelevant phrase (response time [RT]). But because RT may be manipulated by participants, they did not consider it suitable as a measure (Farwell & Donchin, 1991, p. 540). However, other studies using RTs suggest that rapid responses (faster than 800 ms) are not easily affected by intentional manipulation (Posner & Snyder, 1975a, 1975b; Ratcliff & McKoon, 1981). Because the ERP apparatus and data analysis can be costly and complex, RT measures may be preferable. But first one must determine whether RTs alone can reliably detect guilty knowledge and whether they are subject to participants' strategic manipulation.

Our goal in this study was to examine the feasibility of detecting guilty knowledge using only RT measures. Initially, we attempted to replicate the success of Farwell and Donchin's (1991) ERP paradigm using RT alone. Using the same test items, we designed a modified procedure to allow all testing to occur within a single session. In an additional set of experiments, we examined guilty participants' ability to avoid detection by strategically manipulating their RTs.

Experiment 1

The goal of Experiment 1 was to replicate Farwell and Donchin's (1991) paradigm using RTs alone to demonstrate differences between "guilty"[2] and "innocent" test blocks. In particular, we expected participants' mean RTs for phrases related to a crime they committed to be slower and less accurate than RTs for items from a crime to which they had not been exposed.

Our procedure followed the test paradigm used in Farwell and Donchin (1991) with several modifications. First, the crime that participants committed in the present study involved the apparently illicit use of a computer account rather than enacting a spy crime scenario. Second, the participants both committed the crime and completed the recognition task in the same 1-hr experimental session rather than in two sessions with a 24-hr delay. Finally, no ERP measures were collected.

Method

Participants

Thirty-five undergraduates, enrolled in an introductory psychology course at a large midwestern university, received credit for their participation in the study.

Apparatus and Materials

Experimental stimuli were displayed on an IBM-compatible desktop computer with a medium-fast phosphor display and a standard 101-key keyboard. Stimulus control was mediated by the COGSYS[3] (Ratcliff & Layton, 1981) stimulus presentation and response collection program. Participants were seated approximately 2 ft (0.6 m) from the display device.

The test materials were identical to those of Farwell and Donchin (1991; see Appendix). Two complete sets of items were prepared, and the assignment of each set was determined at random. Each set was seen in the guilty test block by half of the participants and in the innocent test block by the other half.

Design and Procedure

As in Farwell and Donchin (1991), the participants completed a series of tasks in the following order: a crime scenario learning task, a crime scenario execution task, an intervening distractor task, and, finally, a phrase classification task.

Scenario learning task. All participants read a cover story that asked them to cooperate with a police investigation during the experiment, specifically to help catch a group of students in a computer crime. The advantage of this procedure is that participants can plausibly commit the crime during the experimental session without leaving the test room. Each participant was asked to log in to a university computer account and send one of the other suspects an electronic message containing incriminating

[2] Use of the terms *guilty* and *innocent* is meant to correspond to terminology in Farwell and Donchin (1991). No suggestion is intended that participants are (or feel) "guilty" or "innocent" as a result of participating in the crime scenario.

[3] COGSYS (Ratcliff & Layton, 1981) has been modified extensively since this formal description. For this experiment, modules were added to handle lists, randomization, extended user text input, and looping structures.

information, signing the message as if he or she were another member of the criminal group.[4]

Initially, participants learned specific information they would use to commit the crime. Training software presented six critical items that participants were to memorize. For example, participants were told that they were to pose as a person named *Phil Jenks* and that they should log in to a computer account under that name. Other critical items included the message that the participant was to send, for example, a note to the alias *Blue Coat* to "bring the *Rain File* containing the *Sub Plans* to *Perch Street*." After presentation of all of the information, participants completed a cued recall task; for example, "username:" was presented as a prompt for "Phil Jenks." The participants typed in an answer at each prompt. This study–test sequence was repeated three times, as in Farwell and Donchin (1991).

Scenario execution task. Next, participants were instructed to execute the instructions they had just learned and actually commit the computer crime. A computer display was presented with what appeared to be an interface to the university computer network. Each participant, following the studied scenario, logged in under the name of an alleged criminal and sent electronic mail to another suspect, including instructions to bring a particular stolen file to a specified location. Although the task appeared realistic to participants, the interface was actually presented via a shell program, and no access to the university system ever occurred. The task ended once participants successfully "sent" the electronic message and logged out.

Distractor task. After the crime had been committed, participants worked on a distractor task for a 10-min period. The task consisted of 30 mathematical word problems (taken from Patalano & Seifert, 1994) designed to engage the participant while preventing any rehearsal of the crime information.

Phrase classification task. After the distractor task, each participant performed an ostensibly unrelated binary classification task. Participants were tested in two blocks: one in which probe items were taken from the crime scenario they had committed earlier (guilty) and another in which probe items were taken from the other crime scenario, which the participant had not seen (innocent). The type of block tested first was determined at random for each participant.

Before each test block, participants learned a set of target phrases they were to recognize and respond to affirmatively when presented as test items. Each of the two target sets consisted of 6 two-word phrases. These target phrases were very similar to the items learned in the earlier crime scenario (as shown in the Appendix). The target items were presented and tested with a recall test three separate times (as described earlier). After training, a classification trial block began, and a randomized series of 108 two-word phrases was presented, one phrase at a time, on the computer screen. The participants' task was to identify target phrases (i.e., words from the list they had just learned) whenever presented and to reject all other phrases. The participants were instructed to press one key (with the right index finger) in response to target phrases and another key (with the left index finger) in response to any other stimulus.

Throughout the phrase classification task, participants were urged to respond as quickly and as accurately as possible. On trials in which the participant's response spanned more than 1,000 ms, the message "Too Slow" was presented on the screen for 1 s before the next stimulus was displayed; otherwise, no feedback was given. The interstimulus interval was randomly varied at 500, 800, or 1,100 ms to prevent response preparation and rhythmic response patterns.

Each of the six target items was repeated three times per block. The remaining 90 trials per block consisted of nontarget items taken from one of two stimulus sets. First, irrelevant items were new phrases that were neither items from the target list nor items from either crime scenario. For each target item, there were four similar, but not identical, irrelevant items, for a total of 24. Each of these items was repeated three times per block, for a total of 72 presentations. The second set of nontarget stimuli was composed of probe items. Within the guilty test block, the six probe items were the critical items from the crime that the participant had committed. For the innocent test block, the six probe items were taken from the second stimulus set, which the participant had not seen before. Each of the six probe items was repeated three times, for a total of 18 presentations per block.

Note that in blocks in which participants were innocent of the crime referred to by the probe items, there were only two categories of stimuli: 17% were target items, and 83% were irrelevant items. For these blocks, probe items were indistinguishable from irrelevant items (all were new to the participants, and the item assignments had been determined at random). However, for blocks in which participants were guilty of the crime referred to by the probe items, there were three categories of stimuli: 17% were target items, 17% were probe items (familiar from the crime), and 66% were irrelevant items. In both types of blocks, accuracy required a response of yes to the 18 presentations of the target phrases and a response of no to the 90 other phrases presented. Thus, a comparison between responses to probe items from guilty blocks and responses to probe items from innocent blocks should demonstrate whether guilty knowledge of the computer crime affected participants' ability to correctly reject those phrases (as nontargets). The entire experimental procedure took approximately 45 min.

Results

An alpha level of .05 was used for all statistical tests.

Exclusions

Two participants were excluded from the analysis because they did not complete the crime (scenario execution task): One participant was too unfamiliar with the computer system, and another refused to perform the task because of concern that the action was "illegal." Six participants were also excluded because they failed to learn at least five of six phrases by the third cued recall test in the scenario learning task of either test block.

Accuracy

Table 1 shows accuracy results for guilty and innocent trials for both nontarget stimulus types. The correct response for both probe and irrelevant items was no. A 2 (guilt) × 2 (stimulus type) repeated measures analysis of variance (ANOVA) performed on accuracy scores revealed a main effect of guilt, $F(1, 28) = 13.76$, $p < .001$, and a Guilt × Stimulus Type interaction, $F(1, 28) = 12.83$, $p < .01$, $\eta^2 = .31$. There was no reliable main effect of stimulus type on the accuracy measure, $F(1, 28) = 1.21$.

The interaction suggests that the accuracy rate for probe items was lower in the guilty condition, whereas accuracy for irrelevant items did not reliably differ as a function of guilt. A contrast ANOVA on the interaction with this pattern was reliable, $F(1, 28) = 6.02$, $p < .02$, $\eta^2 = .18$.

Response Times

The comparison of interest was the RT for probe versus irrelevant items on trials in which participants correctly rejected these

[4] This manipulation was surprisingly effective. Some participants refused to participate in our study because it appeared to be in violation of university computing policy. Others required verbal confirmation that they would not "get in trouble" before sending the e-mail message. In debriefing, many participants indicated that they believed it was "wrong" to send the e-mail message despite being aware that they were taking part in a psychology experiment.

Table 1
Percentages of Correct Rejections for Each Stimulus Type by Experiment

Experiment and condition	Stimulus type: Probe %	Probe SD	Irrelevant %	Irrelevant SD
Experiment 1				
Guilty	74	26	81	11
Innocent	96	19	84	14
Experiment 2a				
Guilty	81	12	85	4
Innocent	99	2	86	4
Experiment 2b				
Guilty	79	15	84	6
Innocent	99	1	87	4

items (answered no, because they were nontarget stimuli). Mean correct RTs to probe and irrelevant stimuli are shown in Figure 1. Responses in the guilty and innocent conditions showed distinct patterns. A 2 (guilt) × 2 (stimulus type) repeated measures ANOVA revealed a main effect of guilt, $F(1, 26) = 84.74$, $p < .0001$; a main effect of stimulus type, $F(1, 26) = 44.61$, $p < .0001$; and a Guilt × Stimulus Type interaction, $F(1, 26) = 56.73$, $p < .0001$, $\eta^2 = .69$.

As expected, mean RTs for probe items were slower in the guilty condition than in the innocent condition. No reliable difference was observed for irrelevant items by condition. A contrast ANOVA on the interaction with this pattern was reliable, $F(1, 26) = 64.90$, $p < .0001$, $\eta^2 = .71$. This interaction pattern, in which the innocent and guilty conditions differed only in regard to a slower mean RT for probe items, characterizes a guilty knowledge effect.

Because probe and irrelevant trials both required responses of no during the recognition task, the magnitudes of the RT effects could be directly compared. In particular, guilty participants required, on average, 300 additional milliseconds to correctly respond to a probe item than to an irrelevant item, whereas innocent participants required no longer to respond to probe than to irrelevant test items. Presumably, this difference reflects the interference of guilty knowledge when participants were attempting to quickly reject probe items related to the crime.

Discussion

Experiment 1 demonstrates a guilty knowledge effect in which both RT and accuracy for probe items revealed participants' knowledge of the crime scenario. To assess the ability of the RT measure to discriminate innocent from guilty blocks, we conducted a discriminant function analysis (Tabachnick & Fidell, 1996). Using the mean difference between correctly rejected probe and irrelevant RTs to predict guilt, this analysis yielded reliable discrimination, $F(1, 53) = 50.19$, $p < .0001$, $\eta^2 = .49$. The resulting function correctly classified guilty participants (i.e., participants' responses during guilty blocks) 89% of the time and innocent participants 100% of the time (overall discrimination accuracy: 95%).

These results are comparable to the ERP analysis performed by Farwell and Donchin (1991), which showed that averaged P300 response patterns for individual participants could be accurately classified 90% of the time for guilty trials and 85% of the time for innocent trials (overall classification accuracy: 87.5%). To obtain these results, Farwell and Donchin (1991) used a complex bootstrapping technique (Efron, 1979; Wasserman & Bockenholt, 1989). Using a somewhat different method, Allen, Iacono, and Danielson (1992) showed that a Bayesian combination of components of the ERP waveform (e.g., P300 amplitude and P300 area) can correctly classify ERPs 94% of the time and that 97% discrimination accuracy can be obtained when a Bayesian combination of ERPs, RT, and accuracy data is used.

We concluded that because RT measures demonstrate success comparable to that of ERPs, and because they are less complicated and less costly, they appear promising as an alternative. However, Farwell and Donchin (1991) claimed that an RT-based measure alone is inaccurate because responses are too easily manipulated by participants. But other studies involving RT measures have demonstrated that rapid responses are not influenced by participants' strategies (Posner & Snyder, 1975a, 1975b; Ratcliff & McKoon, 1981). Ratcliff and McKoon suggested that responses occurring within 800 ms of stimulus onset are unaffected by slow-acting intentional processes. In Experiments 2a and 2b, we addressed participants' ability to manipulate their responses by revealing the intent of the experiment. We informed participants about the probe stimulus category and instructed them to attempt to foil the test. In Experiment 2b, participants were given more detailed information to help them "beat the test."

Experiment 2

Experiments 2a and 2b examined whether participants can manipulate their responses and avoid the detection of critical knowledge. If warned to avoid detection, participants may be able to sufficiently mask the expected differences on guilty-probe trials. For example, suppose naive participants adopt a strategy in which all familiar items are initially accepted as targets, resulting in a

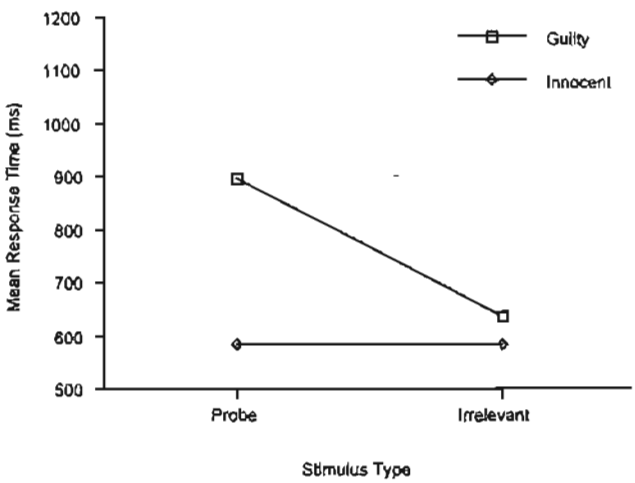

Figure 1. Mean response times for correct rejections for each stimulus type, by condition: Experiment 1.

higher rate of "false-alarm" errors. If participants are warned to expect to see familiar stimuli from the crime, they may use a more cautious strategy for evaluating stimuli. Adopting a higher response criterion overall may assist guilty participants in correctly rejecting probe phrases with a response latency similar to that of irrelevant phrases.

The goal of Experiment 2a was to examine whether such advance warning of the appearance of probe category test items is sufficient for participants to strategically mask their knowledge of the crime. If the guilty knowledge effect is driven primarily by surprise, the warning will allow participants to choose a response strategy that may attenuate the effect. If RT can be altered strategically, as Farwell and Donchin (1991) suggested, the measure will be ineffective for participants attempting to avoid detection.

Experiment 2b was designed to test whether more explicit knowledge of the test would aid participants in appearing innocent during guilty blocks. In addition to informing them about the probe category of the test items, we told Experiment 2b's participants about the expected pattern of results (slower RTs for crime-related words). By preparing participants to try to hide their knowledge, we tested whether the guilty knowledge effect can be attenuated by informed participants.

Method

Participants

Twenty undergraduates, enrolled in an introductory psychology course at a large midwestern university, received credit for their participation in Experiment 2a; 17 participated in Experiment 2b.

Apparatus and Materials

The apparatus and materials in Experiments 2a and 2b were identical to those described for Experiment 1.

Design and Procedure

The design of Experiments 2a and 2b was exactly as in Experiment 1, except that just before the task, participants read either a general or a detailed warning about the task. Specifically, participants in Experiment 2a were warned about the lie-detection intent of the experiment and were told to avoid responding differently to probe and irrelevant items. In Experiment 2b, participants were also told about the expected RT pattern for all conditions and were urged to learn the target list particularly well to facilitate recognizing targets more readily than probes from the crime. They were also urged to use the more cautious strategy of evaluating context (identifying whether it was familiar from the crime or from the target list) instead of the faster strategy based on familiarity alone. We also warned them that, although they should be cautious, they should not respond too slowly (to avoid the 1-s "Too Slow" feedback message). Finally, it was suggested that participants in Experiment 2b would receive fewer experimental trials if they were successful in appearing innocent. This incentive may have been compelling because of the number of trials (108 per block). However, all participants actually completed the same number of trials.

Results

Experiment 2a

Seven participants were excluded from the analysis because they failed to successfully recall five of six items in the scenario learning or phrase classification task. Two other participants were excluded for failing to follow instructions, presumably when attempting to foil the test: One participant responded to all of the target items incorrectly (with responses of no), and another responded to all probe items incorrectly (with responses of yes). In both cases, participants' use of a strategy that violated instructions was readily apparent.

Table 1 shows the accuracy results for innocent and guilty trials by stimulus type. A 2 (guilt) × 2 (stimulus type) repeated measures ANOVA on accuracy revealed a main effect of guilt, $F(1, 10) = 12.75$, $p < .01$, and a main effect of stimulus type, $F(1, 10) = 9.83$, $p < .05$. The Guilt × Stimulus Type interaction was also reliable, $F(1, 10) = 17.18$, $p < .01$, $\eta^2 = .63$. The interaction again suggests that accuracy for probe items was lower in the guilty condition but did not differ in the irrelevant condition.

The pattern of mean RTs for correct responses, shown in Figure 2, reveals the guilty knowledge effect identified in Experiment 1. A 2 (guilt) × 2 (stimulus type) repeated measures ANOVA revealed a main effect of guilt, $F(1, 10) = 53.07$, $p < .0001$; a main effect of stimulus type, $F(1, 10) = 42.45$, $p < .0001$; and a Guilt × Stimulus Type interaction, $F(1, 10) = 50.99$, $p < .0001$, $\eta^2 = .84$. As in Experiment 1, the innocent and guilty conditions differed only in a slower mean RT for probe items, and a contrast ANOVA on the interaction with this pattern was reliable, $F(1, 10) = 53.20$, $p < .0001$, $\eta^2 = .84$.

Experiment 2b

Three participants were excluded from the analysis because they failed to successfully recall five of six critical items in the scenario learning task. Table 1 shows accuracy rates for trials in the innocent and guilty conditions by stimulus type. A 2 (guilt) × 2 (stimulus type) repeated measures ANOVA revealed a main effect of guilt, $F(1, 13) = 19.23$, $p < .001$, and a Guilt × Stimulus Type interaction, $F(1, 13) = 23.76$, $p < .001$, $\eta^2 = .65$, on accuracy. The main effect of stimulus type was not reliable, $F(1, 13) = 3.12$, $p > .10$.

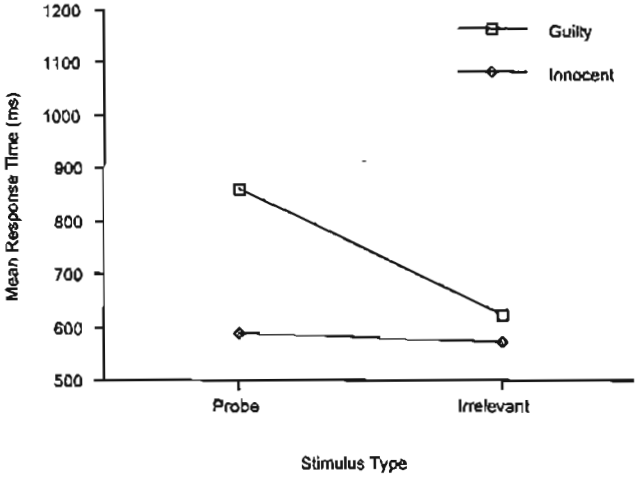

Figure 2. Mean response times for correct rejections for each stimulus type, by condition: Experiment 2a.

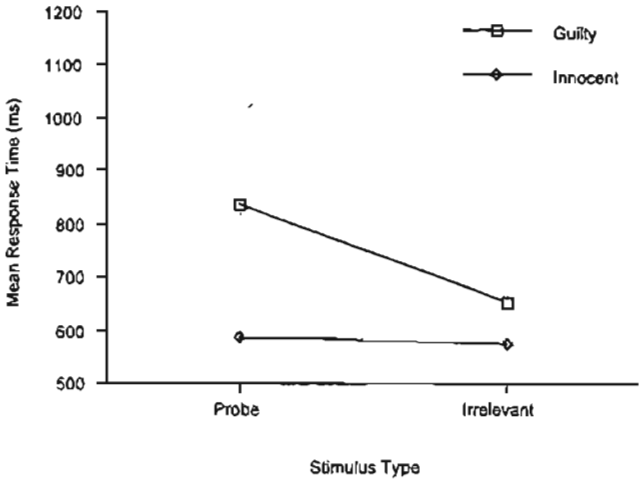

Figure 3. Mean response times for correct rejections for each stimulus type, by condition: Experiment 2b.

RTs for correct responses are shown in Figure 3. A 2 (guilt) × 2 (stimulus type) repeated measures ANOVA revealed a main effect of guilt, $F(1, 13) = 37.70$, $p < .0001$; a main effect of stimulus type, $F(1, 13) = 20.16$, $p < .0001$; and a Guilt × Stimulus Type interaction, $F(1, 13) = 17.33$, $p < .0001$, $\eta^2 = .57$. The same guilty knowledge effect, wherein the only reliable difference as a function of guilt versus innocence involved probe items, was demonstrated with a contrast ANOVA on the interaction with this pattern, $F(1, 13) = 36.26$, $p < .0001$, $\eta^2 = .74$.

Discussion

Experiments 2a and 2b demonstrate that participants' knowledge about the purpose of the phrase classification task does not diminish the utility of RTs in detecting guilty knowledge. Surprisingly, even though participants in Experiment 2a were informed of the special status of the probe items and warned not to respond differently, their data show a pattern strikingly similar to data from Experiment 1. And, in Experiment 2b, despite being fully informed about the task and instructed about how to modify their responses, participants were unable to successfully alter their responses to escape detection. Thus, participants appear to be unable to strategically manipulate their RTs during this task, alleviating concerns raised by Farwell and Donchin (1991). In those few cases in which participants strategically violated the task instructions (e.g., responding yes to all stimuli), their failure to cooperate was easily detected.

Although Experiments 2a and 2b cannot show that no strategy is effective in masking the guilty knowledge effect, they suggest that strategic control of responses may be difficult to achieve, even with detailed knowledge about how the test works. One reason may be the response deadline (the "Too Slow" feedback that appeared 1 s after the stimulus if no response had occurred). Previous studies have shown that RTs shorter than 800 ms may be too quick to allow strategic responses in recognition tasks (Ratcliff & McKoon, 1981). Participants whose RTs are faster than this cutoff may be unable to differentially alter their responses. To consistently respond comfortably before the deadline, participants may be forced to use an automatic strategy (e.g., one based on familiarity or salience). A more conscious strategy may require too much time to execute and therefore fail to conclude within the response window.

Of course, there may exist strategies or practice regimens under which an RT-based guilty knowledge test can be foiled. One possible method may be to rehearse potential irrelevant test items extensively so that they are as familiar as the earlier guilty-probe items. For example, after committing a crime in a red car, one could practice associating the crime with all other possible car colors, in case those items appear on the test. In experimental terms, this expands the proportion of trials with "old" or familiar stimuli, violating the assumptions of the oddball paradigm (Fabiani et al., 1987). Although it would be difficult to conduct such preparation (anticipate the irrelevant test items that may appear), this or other strategies could potentially affect accurate measurement of guilty knowledge. The exact nature of such strategies requires further investigation. The results of Experiments 2a and 2b thus suggest that the guilty knowledge effect, as measured by RT alone, is not easily affected by strategic influences, and thus it can be a valid measure for detecting guilty knowledge.

General Discussion

The experiments reported here investigated whether RTs can reliably detect guilty knowledge. In Experiment 1, using a modified version of an oddball paradigm reported in Farwell and Donchin (1991), we found that RTs were reliably slower and less accurate only when test items were related to a mock crime committed by participants earlier in the session.

Farwell and Donchin's (1991) concern that RT may be subject to strategic manipulation was addressed in Experiments 2a and 2b. In Experiment 2a, we warned participants to expect crime-related probe items and to avoid responding differently to them to mask their knowledge of the crime. In Experiment 2b, we gave participants even more detailed information about the test, along with suggestions about how to appear innocent during guilty blocks. Participants in both Experiment 2a and Experiment 2b were unable to mitigate the guilty knowledge effect in their RTs. We conclude from these experiments that RT is not only a valid and reliable measure of guilty knowledge but also one not easily altered by strategic manipulation.

Analyses suggest that RTs are at least as sensitive as ERPs in this interference paradigm. A discriminant function analysis on the RT data collected in Experiment 1 yielded an overall classification accuracy of 95%, which is comparable to the 87% achieved by Farwell and Donchin (1991) using a bootstrapping method and the 97% achieved by Allen et al.'s (1992) Bayesian classification method. In comparison, polygraph participants who attempt to raise their level of excitation during baseline questioning can render classification accuracy at or below chance (Lykken, 1998).

One limitation of the interference paradigm in Farwell and Donchin (1991) is the method for detecting guilt with an individual test participant. By collecting responses from multiple test participants on the same materials in both conditions, Farwell and Donchin (1991) established cutoffs in the distributions that can reliably determine whether an individual participant's ERPs suggest guilt. However, if the response characteristics for a new set of test stimuli differ from previously tested stimuli (e.g., the test items

are longer or more similar to each other), the previously established cutoffs to differentiate innocence from guilt may be inappropriate. Ideally, one would like to draw a conclusion about a single participant across a wide variety of stimuli without having to collect distributions of responses for the stimulus items from many other participants.

To address this issue, we created an analysis method using data from a single participant in two trial blocks, one in which the participant is known to be innocent and one in which the participant's guilt is to be determined (the "guilty?" condition). The innocent and guilty? blocks are compared in three steps: a test for differences in error rates, a test for differences in RT distributions, and a test for differences in RT variance. A significant difference on any one of the three tests indicates a low probability that the innocent and guilty? response samples were drawn from the same underlying population of responses. Specifically, a Fisher exact (2×2) test compared the guilty? probe error rate with the innocent probe error rate; there was a one-tailed prediction that the guilty? probe error rate would be higher. The Kolmogorov–Smirnov two-sample test (Stephens, Kotz, Johnson, & Read, 1983) was used to compare the guilty? probe RT distribution with the guilty? irrelevant, innocent probe, and innocent irrelevant pooled RT distributions (all of which were new items for participants). The one-tailed prediction was that the guilty? probe RTs would be longer, on average, than the pooled RTs. Finally, an F test for variances was used to compare the variability of the guilty? probe RT sample against the variability of the pooled RT[5] sample. The one-tailed prediction was that the guilty? probe RTs would have higher variance than the pooled RTs. This procedure was tested on the data from Experiments 1 and 2 and resulted in highly accurate classifications of each individual participant at the .05 and .01 criterion levels.

The results showed a .98 hit rate at the .05 significance level and a .93 hit rate at the .01 significance level. Theoretically, this analysis sets the false alarm rate (classifying "innocent" participants as "guilty") at .01 or .05; however, because data from both blocks were pooled in the analysis, there was no way to check this assumption. When the guilty? and innocent conditions were analyzed separately, the innocent-only data provided a direct estimate of the empirical false alarm rate. When the probe and irrelevant distributions were tested separately for each block, the empirical (observed) false alarm rate, using the .05 criterion level, was .02; specifically, one of the 45 trial blocks was classified as guilty when it was actually innocent, and one of the 45 guilty blocks was classified as innocent. Alternatively, using a criterion level of .01 led to an observed false alarm rate of .0.

Several other factors may limit the feasibility of RT measures for guilty knowledge detection. A remaining question is whether longer delays between exposure to crime information and testing will affect results. Because a test may be given long after the occurrence of a crime, it may be essential that the test clearly bring the crime context to mind before interference from crime-related knowledge occurs. Potentially, knowledge of the test's detection purpose may help to reactivate the crime context if it exists in memory. Greene, Gerrig, McKoon, and Ratcliff (1994) showed that reference cues may be sufficient to invoke a prior context during reading comprehension; thus, reference to the crime under investigation may be sufficient to activate any related knowledge a participant possesses. Farwell and Donchin (1991) did report a successful test of 2 participants who had committed actual crimes using ERPs during the interference paradigm.

Two other factors potentially affecting test outcomes are the distinctiveness and similarity of test items. The probe items must be distinctive to access the correct referent in memory. For example, if the guilty-probe items refer to generic concepts (e.g., "street" or "file"), access to a specific crime episode in memory ("Perch Street") may not occur. Other stimulus variables such as word frequency and length may introduce differences between the stimuli in the guilty and innocent conditions, compromising interpretation of the test results. In these experiments, all items were equated and then assigned at random to appear in the crime. However, with an actual crime providing probe items, randomization is not possible, and great care would be needed to equate probe items with target and irrelevant items.

From this research, we conclude that RT measures of guilty knowledge are a viable alternative to ERP and polygraph methods. Critical elements of the paradigm include the oddball procedure, adequate learning of probe (crime) and target items, accurate measurement of RT and accuracy, and a short response deadline to avoid potential strategic manipulation by participants. The experiments reported here provide evidence that RTs are at least as accurate as, and more reliable and less subject to conscious manipulation than, other methods. Because the apparatus required is fairly inexpensive (a software program running on a DOS-based computer), RT-based tests may have an advantage in cost and ease of use, and, with appropriate analyses, they can deliver an immediate and unambiguous classification with high accuracy.

[5] The significance levels of the Kolmogorov–Smirnov (Stephens, Kotz, Johnson, & Read, 1983) and variance-ratio tests were determined by a randomization procedure with 1,000 pseudo-random trials. A Bonferroni correction was used to keep the subjectwise significance level at its nominal value of .05..

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., & Lieblich, I. (1986). Trial by polygraph: Scientific and juridical issues in lie detection. *Behavioral Sciences and the Law, 4,* 459–479.

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

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*(4), 19–33.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11,* 357–374.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

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.

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

Furedy, J. J. (1991). Alice in Wonderland terminology usage in, and communicational concerns about, that peculiarly American flight of technological fancy: The CQT polygraph. *Integrative Physiological and Behavioral Science, 26,* 241–247.

Furedy, J. J., & Heslegrave, R. J. (1988). Validity of the lie detector: A psychophysiological perspective. *Criminal Justice and Behavior, 15,* 219–246.

Greene, S. B., Gerrig, R. J., McKoon, G., & Ratcliff, R. (1994). Unheralded pronouns and management by common ground. *Journal of Memory and Language, 33,* 511–526.

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

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: McGraw-Hill.

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

Posner, M. I., & Snyder, C. R. (1975a). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 145–175). Hillsdale, NJ: Erlbaum.

Posner, M. I., & Snyder, C. R. (1975b). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt (Ed.), *Attention and performance* (Vol. 5, pp. 669–682). London: Academic Press.

Ratcliff, R., & Layton, W. M. (1981). A microcomputer interface for control of real-time experiments in cognitive psychology. *Behavior Research Methods & Instrumentation, 13,* 216–220.

Ratcliff, R., & McKoon, G. (1981). Automatic and strategic priming in recognition. *Journal of Verbal Learning and Verbal Behavior, 20,* 204–215.

Rosenfeld, J. P. (1995). Alternative views of Bashore and Rapp's (1993) "Alternatives to traditional polygraphy": A critique. *Psychological Bulletin, 117,* 159–166.

Rosenfeld, J. P., & Bessinger, G. T. (1990). Feedback-evoked P300 responses in lie detection. *Psychophysiology, 27*(Suppl. 4A), 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.

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

Stephens, M. A., 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). New York: Wiley.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper & Row.

United States v. Scheffer, 44 M. J. 443 (1998).

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology, 26,* 208–221.

Appendix

Test Materials Used in Present Study and Farwell and Donchin (1991)

Probe	Target	Irrelevant	Probe	Target	Irrelevant
Blue Coat	Green Hat	Brown Shoes	White Shirt	Green Tie	Beige Suit
		Red Scarf			Red Vest
		Gray Pants			Tan Belt
		Black Gloves			Black Socks
Phil Jenks	Tim Howe	Ray Snell	Dale Spence	Wayne Bryant	Glenn Platt
		Neil Rand			Walt Rusk
		Gene Falk			Tod Ames
		Ralph Croft			Earl Dade
Op Cow	Op Pig	Op Horse	Op Spruce	Op Fir	Op Oak
		Op Goat			Op Birch
		Op Sheep			Op Elm
		Op Mule			Op Pine
Rain File	Snow File	Hail File	Owl File	Swan File	Wren File
		Wind File			Duck File
		Sleet File			Crow File
		Fog File			Goose File
Sub Plans	Ship Plans	Tank Plans	Brass Plans	Steel Plans	Tin Plans
		Plane Plans			Zinc Plans
		Bomb Plans			Lead Plans
		Gun Plans			Iron Plans
Perch Street	Shark Street	Cod Street	Lion Street	Fox Street	Deer Street
		Carp Street			Wolf Street
		Pike Street			Bear Street
		Trout Street			Elk Street

Note. Op = operation.

Received February 5, 1998
Revision received February 22, 1999
Accepted February 26, 1999 ■