Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: A reexamination

Gershon Ben-Shakhar[1] and Eitan Elaad[2]

1. Department of Psychology, The Hebrew University of Jerusalem

2. Division of Identification and Forensic Science, Israel National Police H.Q

Running Head: Items' repetition in the GKT

Authors' Notes

We wish to thank Limor Bar, Neta Bar, Sossy Fuchs, Michal Maimaran and Erga Sinai for their help in the data collection and analysis.

Address requests for reprints to:

Professor Gershon Ben-Shakhar

Department of Psychology

The Hebrew University of Jerusalem

Jerusalem 91905 Israel

ABSTRACT

The effect of question repetition and variation on the efficiency of the Guilty Knowledge Test (GKT) was examined in a between-subjects experiment with 3 conditions. Each participant was presented with a sequence of 12 biographical questions (e.g., father's name, month of birth). In Condition '1', a single question was repeated 12 times; in Condition '4', each of four different questions was repeated 3 times; and in Condition '12', twelve different questions were used. In each condition, skin conductance and respiration responses to the relevant items were used to differentiate between "Guilty" (individuals whose true biographical items were presented) and "Innocents" (individuals whose true biographical items were not presented). A monotonic relationship between the number of different questions used and detection efficiency was observed only with the electrodermal measure (the areas under the ROC curves, obtained with this measure in conditions '1', '4', and '12' were .68, .81, and .99, respectively). The results of this study demonstrate that a GKT based on multiple questions is superior to the use of many repetitions of a single or a few questions, and it can reach an almost perfect detection efficiency.

Key Words: Guilty Knowledge Test; Polygraph; Psychophysiological detection; Respiration changes; Skin conductance responses

Effects of questions' repetition on the efficiency of the guilty knowledge test: A reexamination

The study of psychophysiological detection has attracted a great deal of research attention and has become an important area of applied psychology (e.g., Ben-Shakhar & Furedy, 1990; Lykken, 1998; Raskin, 1989). In this study we focus on just one of the two prominent methods of psychophysiological detection, known as the Guilty Knowledge Test (GKT) or the Concealed Information Test (CIT). This method is based on sound theoretical principles and proper controls and therefore it satisfies the necessary requirements of an objective test (see, Ben-Shakhar, Bar-Hillel & Kremnitzer, 2001; Ben-Shakhar & Elaad, 2001a; Lykken, 1974, 1998).

The GKT (Lykken, 1959, 1960) utilizes a series of multiple-choice questions, each having one relevant alternative (e.g., a feature of the crime under investigation) and several neutral (control) alternatives, chosen so that an innocent suspect would not be able to discriminate them from the relevant alternative (Lykken, 1998). Typically, if the suspect's physiological responses to the relevant alternative are consistently larger than to the neutral alternatives, knowledge about the event (e.g., crime) is inferred. As long as information about the event has not leaked out, the probability that an innocent suspect would show consistently larger responsivity to the relevant than to the neutral alternatives depends only on the number of questions and the number of alternative answers per question, and hence it can be controlled such that maximal protection for the innocent is provided.

Extensive research conducted since the early Sixties has demonstrated that the GKT can be successfully used for detecting relevant information and discriminating between knowledgeable (guilty) and innocent individuals (e.g., Ben-Shakhar & Furedy, 1990; Elaad, 1998; Lykken, 1959, 1960, 1998). Recently, Ben-Shakhar and Elaad (2001b) conducted a

meta analysis of GKT research and showed that when properly administrated the GKT can reach an average correlation coefficient as high as 0.71 between the detection measure and the criterion of guilt versus innocence.

A successful implementation of the GKT depends on two conditions that must be satisfied:

1. Salient features of the event must be identified. The GKT rests on the assumption that these features are perceived by individuals exposed to the event in question, and remembered at the time of the investigation.
2. These features must be concealed from the general public, and must not be leaked to potential examinees.

However, in spite of the extensive research supporting its validity, the GKT is not being used very often in actual investigations. Podlesny (1993) conducted a survey of FBI polygraph investigations and estimated that the GKT might have been used in only 13.1% of them. This low estimate can be explained by the difficulty involved in identifying several proper GKT questions (a proper GKT question is a specific case fact that is very likely to be known to a guilty, but not to an innocent person). However, Podlesney's estimate must be treated cautiously because FBI investigators are probably not motivated and trained to search for features of the crimes that can be used to formulate GKT questions. It is generally assumed that at least 4 different questions are required to construct a GKT investigation, and even if Podlesney underestimated the number of cases in which the GKT might have been used, there are many cases where it is difficult to construct 4 different proper GKT questions. Several different questions are needed because innocent suspects may display enhanced responses to the relevant item of one or two questions just by chance. On the other hand, a consistent pattern of enhanced responses to the relevant items of several GKT questions is much less likely if the examinee has no knowledge of the event under investigation.

Recently, Elaad & Ben-Shakhar (1997) reported two mock crime experiments, in which similar detection efficiencies were obtained with a single question repeated 12 times, and with 4 different questions repeated 3 times. If this finding can be generalized, it would mean that the GKT could be much more applicable than previously believed. If differential physiological responsivity is relatively resistant to habituation, it might be possible to obtain satisfactory correct detection rates with many repetitions of a few or even a single GKT question, provided that several physiological measures are being used.

The present study was designed as a constructive replication of Elaad and Ben-Shakhar's (1997) study. Instead of a mock crime procedure, we used personal information items (e.g., first name, mother's name, month of birth). This procedure has been used in several previous studies (e.g., Ben-Shakhar, Lieblich, & Kugelmass, 1970; Elaad & Ben-Shakhar, 1989) and produced high detection efficiency. In addition, we compared three, rather than two experimental conditions: (1) A single question repeated 12 times (Condition '1'). (2) Four different questions repeated 3 times (condition'4'). (3) Twelve different questions, each presented once (condition'12'). As in the previous study, the electrodermal and respiration measures were used, separately and as a combined detection measure. Unlike the previous study, motivational instructions were given to all participants. Finally, the order of presentation of the GKT questions was changed (for a detailed description see below).

METHOD

Participants – A total of 108 undergraduate students (77 women and 31 men) participated in the experiment for payment or course credit. Their mean age was 22.3 years (SD = 2.7 years).

Apparatus - Skin conductance was measured by a constant voltage system (0.5 V Atlas Researches). Two Ag/AgCl electrodes (0.8 cm diameter) were used with an electrode paste which was prepared according to the recipe provided by Fowles et al. (1981). Respiration was recorded by a pneumatic tube positioned around the thoracic area. The experiment was

conducted in an air-conditioned laboratory, and was monitored from a control room separated from the laboratory by a one-way mirror. A Macintosh II computer was used to control the stimulus presentation and to compute skin conductance and respiration changes. The stimuli were displayed on a Macintosh 13" color monitor.

Design – A 3 by 2 between-participants design was used, with the following 2 factors: (1) Number of questions used (1, 4, and 12); (2) Guilt condition ("guilty participants" whose personal information items were included in the stimulus sequence, and "innocent participants" whose personal items were not presented). Twenty four participants were allocated to each of the 3 "guilty" conditions, and 12 were allocated to each of the "innocent" conditions.

Stimuli – Each participant filled a 12-questions questionnaire (e.g., first name, mother's name, place of birth). The answers to these questions served as the relevant items for the "guilty" participants. Control, items were chosen for each participant, such that they differed from the relevant items, but were of the same category. In conditions '1' and '4', where not all 12 questions were used, the questions were chosen such that each question's category (e.g., first name) was used with the a fixed frequency. For example, in condition '1', where a single question was repeated 12 times, the first name was used as the single GKT question for 2 guilty and one innocent participant.

Procedure - The experiment was conducted in two sessions separated by 2-3 days. In the first, participants filled the personal-items questionnaires, and in the second the GKT was administered. An experimenter who was unaware of the guilt condition to which the examinee was assigned, attached the polygraph devices, and conducted the GKT examination. Participants were told that the experiment was designed to test whether they could cope with the polygraph test and avoid detection of their personal items. They were promised a bonus of 5 NIS (about $1.25 at the time of the study) for a successful performance of the task. All

examinees were presented with 12 GKT questions, arranged in 3 blocks of 4 questions. In condition ‘1’, a single GKT question was repeated 12 times; in condition ‘4’, each of four different questions was repeated 3 times, such that each block contained all 4 different questions; in condition ‘12’, each of 12 different questions was presented just once. The pre-recorded questions were presented through the computer. Following the presentation of each question, a buffer item was presented, followed by five alternative answers. Examinees were requested to respond "no" to every item, thus denying knowledge of the relevant item. The order of the five items was randomly determined and the ISIs ranged from 16 to 24 s, with a mean of 20 s. A 1 min. rest period was allowed after presentation of each block of 4 questions. At the end of the experiment participants were debriefed and paid.

Response Scoring and Analysis

(a) Electrodermal responses: Examinees' responses were transmitted in real time to the Macintosh II computer. The maximal conductance change obtained from the examinee, from 1 s to 5 s after stimulus onset was computed using an A/D (NB-MIO-16) converter with a sampling rate of 1000 per second. To eliminate individual differences in responsivity and permit a meaningful summation of the responses of different participants, each examinee's conductance changes were transformed into within-subjects standard scores (Ben-Shakhar, 1985). The Z scores used in this study were computed relative to the mean and standard deviation of the examinee's response distribution within each block of 4 questions (24 items). Within-blocks Z scores were used because they are more resistant to habituation effects, and therefore more efficient (Elaad & Ben-Shakhar, 1997).

(b) Respiration: The respiration responses were defined on the basis of the total respiration line length (RLL) during the 15 s interval following stimulus onset. Timm (1982) noted that the computation of the RLL from the curvilinear respiration pattern might be disproportionally affected by the starting point of measurement. For example, starting from a

point in the rapidly ascending inspiration curve, and from a point at the end of the expiration curve, where changes are relatively slow, would produce different RLLs for equal time intervals. To deal with this problem, we followed the procedure used by Elaad, Ginton and Jungman (1992) and defined each response as the mean of ten length measures (0.1 s. after stimulus onset through 15.1 s after stimulus onset, 0.2 s through 15.2 s after stimulus onset, etc.). In other words, ten 15-seconds windows were created, each beginning 0.1 s later than the previous window, and the RLL was defined as the mean of the ten length computed for the ten windows. Each RLL was computed using a sampling rate of 20 per second. Similar standardization transformation was applied for the RLL as the one described above in relation to the electrodermal measure. But since guilty knowledge is reflected by smaller rather than larger RLLs, the RLL Z scores were multiplied by –1 and are presented as positive values in all subsequent analyses.

A combined measure was defined as a sum of the SCR and RLL Z scores. Due to mechanical problems with the respiration-measurement equipment, the RLL data of 11 guilty participants were lost. Thus, all the analyses of the RLL and the combined measure are based on restricted sample sizes.

The personal items served as the relevant stimuli for the "guilty" examinees and a randomly predetermined item in each question served as the relevant item for the "innocent" examinees. The standardized responses to these items were used as the dependent variables in all subsequent statistical analyses. A rejection region of $p<.05$ was used for all statistical tests.

RESULTS

The Z scores corresponding to the relevant stimuli were averaged across questions and across participants within each experimental condition and each block of questions, as well as across blocks. A difference score was computed for each measure within each block, as the mean Z score of the “guilty” participants minus the mean Z score of the “innocent”

participants. These difference scores are presented in Table 1 as a function of blocks and experimental conditions, as well as across blocks, for each physiological measure. In addition, the means and standard deviations of the standardized responses computed across blocks for each experimental condition and each physiological measure are displayed in Table 2.

Insert Tables 1 and 2 about here

A three-way ANOVA was conducted for each measure (SCR, RLL and the combined measure), with two between-subjects factors: Guilt status ("Guilty" vs. "Innocent" participants); Experimental Condition (‘1’, ‘4’, or ‘12’), and one within-participants factor (questions’ block) with three levels. The results revealed a statistically significant main effect for Guilt status, with each measure [$F(1,102)=44.01$, $MSE=0.51$ for the SCR; $F(1,91)=34.70$, $MSE=0.34$ for the RLL; and $F(1,91)=59.03$, $MSE=1.08$ for the combined measure]. The proportions of the total variance of the three dependent measures accounted for by the main effect of guilt vs. innocence were, 27.2%, 25.6% and 34.8% for the SCR, RLL and the combined measure, respectively. The 2-way interaction (Guilt by Experimental Condition) produced statistically significant effects with the SCR and the combined measure [$F(2,102)=4.70$; $F(2,91)=3.90$, respectively], but not with the RLL [$F(2,91)=0.37$]. These interactions, which account for 5.8% and 4.6% of the variance of the SCR and the combined measure, respectively, indicate that the Z score differences between "guilty" and “innocent” participants tend to increase with an increase in the number of different questions. The blocks factor produced a statistically significant main effect only for the SCR [$F(1,102)=4.30$, $MSE=0.20$, $\varepsilon=0.98$], but it did not interact with any of the between-participants factors. No significant main or interaction effects were obtained for the blocks factor with the other measures.

In addition to group data, it is interesting to look at classification accuracy of individual examinees. To achieve this goal, the Lykken’s scoring procedure was adopted

(Lykken, 1959). By this procedure, the standardized responses, of each physiological measure, to all alternatives of each question are rank-ordered. If the relevant alternative elicits the largest response, a value of 2 is assigned to the question, if it elicits the second largest response, a value of 1 is assigned to the question, otherwise, a value of 0 is assigned. These values are then summed up across all 12 questions (or repetitions) to produce a single detection score (ranging between 0 and 24) for each participant on each physiological measure (SCR, RLL and their combination). A cutoff score of 12 was set on each of these detection measures (a detection score of at least 12 yielded a "guilt" classification). Rates of correct classifications of guilty and innocent participants based on each measure are presented in Table 3 as a function of experimental conditions.

Insert Table 3 about here

The accuracy rates presented in Table 3 depend on a single, arbitrary cutoff point. Furthermore, it is difficult to compare experimental conditions on the basis of accuracy rates because often gains in one type of classification (e.g., classification of "guilty" examinees) is associated with losses in the other classification category. For example, an inspection of the SCR classification rates (see Table 3) demonstrates that, while condition '12' is clearly superior to the two other conditions, a comparison of conditions '1' and '4' is difficult (condition '4' yields a better classification rate of "guilty" participants, but a worse rate of classifying "innocent" participants, relative to condition '1').

Therefore, an additional approach for describing and comparing detection efficiency was adopted from Signal Detection Theory (e.g., Green & Swets, 1966; Swets, Tanner & Birdsall, 1961). A statistic describing detection efficiency by comparing the entire distributions of Z scores to the relevant alternatives of innocent and guilty examinees was computed for each measure within each experimental condition. On the basis of these distributions, a Receiver Operating Characteristic (ROC) curve was generated for each

experimental condition and each physiological measure (for a detailed description of the ROC construction, see Lieblich, Kugelmass & Ben-Shakhar, 1970). Figure 1 describes the ROC curves, computed on the basis of the combined measure, for the three experimental conditions. In addition, the area under each ROC curve, along with the 90% confidence interval for the area (see, Bamber, 1975), was computed for each measure under each experimental condition. These results, displayed in Table 4, reveal that variations in the GKT questions affected detection efficiency with the electrodermal and the combined measure, but not with the RLL. With the combined measure, the ROC area increased from about 0.80 with 12 repetitions of a single question to an almost perfect detection with 12 different questions.

Insert Figure 1 and Table 4 about hereDISCUSSION

In contrast to the results reported by Elaad and Ben-Shakhar (1997), the present results demonstrate a clear tendency for larger SCR differentiation between guilty and innocent participants with an increase in questions' variation (the SCR areas increased from 0.68 in the single-question condition to 0.81 and 0.99 in conditions '4' and '12', respectively). The results obtained with 12 different questions are particularly impressive and demonstrate an almost perfect detection efficiency (an area of 0.99). This condition was not included in our previous study, but the results of the present study reveal a clear advantage in electrodermal detection efficiency with 4 different questions over the use of a single GKT question.

The areas under the ROC curves can be translated into more conventional measures of effect size. Specifically, if it is assumed that the detection measure (e.g., the Lykken score in our case) has a Normal distribution with a fixed variance for both the "guilty" and the "innocent" groups, then the distance (in standard deviation units) between the means of these two distributions (d') can be derived from the ROC area (for a more detailed description of this derivation, see Ben-Shakhar & Elaad, 2001b; Ben-Shakhar, Lieblich & Bar-Hillel, 1982). The d' values for the SCR measure under conditions '1', '4' and '12' are 0.65, 1.26 and 3.11,

respectively. Adopting Cohen's (1988) criteria (according to which a d' value of 0.80 represents a large effect size), it can be concluded that while a repetition of a single GKT question results in a medium effect size, the use of 4 different questions definitely produces a large effect size. Relying on as many as 12 different questions produces an effect size almost four times larger than what Cohen considered a "large effect size".

This study joins several previous studies, which demonstrated the utility of the RLL measure in psychophysiological detection (e.g., Ben-Shakhar & Dolev, 1996; Elaad, & Ben-Shakhar, 1997; Elaad, et al., 1992; Timm, 1982, 1987). Inspection of Tables 3 and 4 reveals that a combination of the two physiological measures (SCR and RLL) increases detection efficiency beyond that obtained with the electrodermal measure alone.

Interestingly, the respiration measure was unaffected by variations in the questions and in this respect the present results are consistent with those reported by Elaad and Ben-Shakhar (1997). The present study demonstrates that 12 repetitions of a single GKT question can produce a differentiation between guilty and innocent participants with the RLL measure (a ROC area of 0.85, or a d' of 1.46). It should however be noted that inspection of the RLL-based correct classification rates (see Table 3) reveals a monotonic increase in correct detection rates of guilty participants from condition '1' to '12', with a perfect detection of innocent participants in all 3 conditions. As indicated earlier, the correct detection rates reflect a single cutoff point and the true positive rate may increase with slight changes in the cutoff point, with little or no increase in false-positive rates.

Several previous studies suggested that the RLL measure is more robust than the electrodermal measure against various manipulations. For example, two studies reported that mental countermeasures affected psychophysiological detection with the electrodermal, but not with the respiration measure (Ben-Shakhar & Dolev, 1996; Honts, Devitt, Winbush, & Kircher, 1996). Ben-Shakhar, Gronau and Elaad (1999) demonstrated that knowledge of the

relevant information by innocent participants resulted in an increased rate of false-positive outcomes with the electrodermal measure, but not with the RLL. Our results join these studies in indicating that the RLL may be more resistant to habituation than the electrodermal measure (only the SCR showed a significant decline in response magnitude as a function of questions' blocks).

Thus, habituation of the SCRs with repetitions might account for the relatively poor SCR detection efficiency with many repetitions of a single question. However, this finding must be qualified because the blocks effect emerged only as a main effect and did not interact with the guilt factor, or with the questions' variation condition. Thus, if a similar reduction in response magnitude occurs for both guilty and innocents, then the differentiation between these two groups might be unaffected by repetition. An inspection of Table 1 reveals that the differences between "guilty" and "innocents" in relative responding to the critical questions did not show a clear and consistent reduction under any experimental condition and with neither physiological measure. In particular, an inspection of condition '1', which is most vulnerable to habituation, reveals that the SCR difference increased from 0.26 in the first block to 0.38 in the second and then dropped to 0.17 in the third. The small SCR differentiation in the 3rd block of condition '1' may account for the relatively poor performance of the SCR in this condition. The RLL differences, on the other hand, are 0.14, 0.54 and 0.52 in the 3 blocks, respectively and the combined measure differences are 0.38, 0.87 and 0.67 in the 3 blocks, respectively. Thus, it seems that differential responsivity with both measures is relatively resistant to habituation. This "resistance to habituation" phenomenon may be attributed to the use of within-blocks standard scores (see Elaad & Ben-Shakhar, 1997).

From an applied perspective, the results obtained with the combined measure are of the greatest relevance because realistic applications of psychophysiological detection rely on

several physiological measures. When the combined measure is considered, there is a clear advantage for questions' variation because detection efficiency increases monotonically with more variation (the ROC areas obtained under the '1', '4' and '12' conditions are 0.79, 0.87 and 0.99, respectively, which translates into d' values of 1.16, 1.60 and 3.41, respectively). It can be argued that the advantage of using a variety of questions over repetitions of a single question is reflected by the true positive rate (sensitivity) and not by the true negative rate (specificity), because unless a question has been leaked out, there is no reason why an innocent suspect would show a consistent pattern of responding to the correct item even if the same question is repeated over and over again. The case of guilty suspects may be entirely different because one can never be certain that all the details of the event under consideration were perceived during the event and are remembered during the investigation. Thus, if many different questions are used and a guilty suspect failed to notice or remember some of them, he or she can still be detected by consistently responding to the correct items of the other questions[1]. But if a single question is used and the correct item of that question was not noticed, detection would not be possible. This argument is consistent with the present results (see Table 3). At least with the Lykken scoring technique and the cutoff point used, there are negligible differences between conditions '1', '4' and '12' in true negative rates, which are very high under all conditions and with both physiological measures. The true positive rates, on the other hand, increase monotonically from the single- to the twelve-question conditions.

Thus, the conclusion that must be drawn from this study is that efforts should be made to increase the number of GKT questions. It is however, doubtful whether it would be possible to generate as many as 12 different proper GKT questions, even with increased efforts. Thus, from a practical perspective the question is whether the GKT should be used when only 4 different GKT questions, or even less are available. As indicated above, our results show that the effect size produced by 3 repetitions of 4 different questions is about 1.60, which is twice

as large as what Cohen (1988) considered to be “a large effect size”. With 12 repetitions of a single question the effect size estimate decreased to a value of 1.16, which is still a large effect size. So our recommendation would be that the GKT should be used even when only a small number of proper questions was generated. This usage should be based on several repetitions of each question, on several physiological measures and on a within-blocks standardization technique. Of course investigators should be aware of the fact that such a usage may produce a relatively large rate of false negatives, but at least innocent suspects are protected. In addition, it should be remembered that other Psychophysiological detection methods are much more questionable and may result in much greater risks of implicating innocent suspects.

REFERENCES

Bamber, D. (1975). The area under the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, 12, 378-415.

Ben-Shakhar, G. (1985). Standardization within individuals: A simple method to neutralize individual differences in psychophysiological responsivity. Psychophysiology, 22, 292-299.

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2001). Trial by polygraph: Reconsidering the use of the GKT in court. Manuscript in preparation.

Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the Guilty Knowledge Technique: The effects of mental countermeasures. Journal of Applied Psychology, 81, 273-281.

Ben-Shakhar, G., & Elaad, E. (2001a). The Guilty Knowledge Test (GKT) as an application of psychophysiology: Future prospects and obstacles. In: M. Kleiner (Ed.). Handbook of Polygraphy, Academic Press, In press.

Ben-Shakhar, G., & Elaad, E. (2001b).The validity of psychophysiological detection of deception with the Guilty Knowledge Test: A meta-analytic review. Manuscript in preparation.

Ben-Shakhar, G., & Furedy, J. J. (1990). Theories and applications in the detection of deception: A psychophysiological and international perspective. New York: Springer-Verlag.

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. Journal of Applied Psychology, 84,651-660.

Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, M. (1982). An evaluation of Polygrapher's judgments: A review from a decision theoretic perspective. Journal of Applied Psychology, 67, 701-713.

Ben-Shakhar, G., Lieblich, I., & Kugelmass, S. (1970). Guilty knowledge technique: of Application of signal detection measures. Journal of Applied Psychology, 54, 409-413.

Cohen, J.E. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Elaad, E. (1998). The challenge of the concealed knowledge polygraph test. Expert Evidence, 6, 161-187.

Elaad, E., & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection of information. Psychophysiology, 26, 442-451.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of Item Repetitions and Variations on the Efficiency of the Guilty Knowledge Test. Psychophysiology, 34, 587-596.

Elaad, E., Ginton, A. & Jungman N. (1992). Detection measures in real-life criminal guilty knowledge tests. Journal of Applied Psychology, 77, 757-767.

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., & Venables, P. H. (1981). Publication recommendations for electrodermal measurements. Psychophysiology, 18, 232-239.

Green, D.M., & Swets, J.A. (1966). Signal detection theory and Psychophysics. New York: John Wiley & Sons.

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.

Lieblich, I., Kugelmass, S., & Ben Shakhar, G. (1970). Efficiency of GSR detection of information as a function of stimulus set size. Psychophysiology, 6, 601-608.

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. (1974). Psychology and the lie detector industry. American Psychologist, 29, 725-739.

Lykken, D.T. (1998). A Tremor in the Blood: Uses and Abuses of the Lie Detector. New York: Plenum Trade.

Podlesny, J.A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations? A review of FBI case records. Crime Laboratory Digest, 20, 57-61.

Raskin, D.C. (1989). Polygraph techniques for the detection of deception. In D.C. Raskin (Ed.) Psychological methods in criminal investigation and evidence. New York, Springer.

Swets, J. A., Tanner, W.P., Jr., & Birdsall, T.C. (1961). Decision processes in perception. Psychological Review, 68, 301-340.

Timm, H.W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. Journal of Applied Psychology, 67, 391-400.

Timm, H.W. (1987). Effect of Biofeedback on the detection of deception. Journal of Forensic Sciences, 32, 736-746.

Footnotes

1. We thank David Lykken for raising this argument.

Table 1: Mean Z score differences between responses to critical items of "guilty" and "innocent" participants as a function of blocks and experimental conditions for each physiological measure

SCR

	Block 1	Block 2	Block 3	Across Blocks
Single question repeated twelve times	0.26	0.38	0.17	0.27
Four questions repeated three times	0.59	0.44	0.50	0.51
Twelve different questions	0.77	1.16	0.77	0.90

RLL

	Block 1	Block 2	Block 3	Across Blocks
Single question repeated twelve times	0.14	0.54	0.52	0.40
Four questions repeated three times	0.37	0.36	0.31	0.35
Twelve different questions	0.49	0.46	0.53	0.49

SCR + RLL

	Block 1	Block 2	Block 3	Across Blocks
Single question repeated twelve times	0.38	0.87	0.67	0.64
Four questions repeated three times	0.97	0.75	0.69	0.80
Twelve different questions	1.32	1.69	1.37	1.46

Table 2: Means (SDs) of each Physiological Measure as a Function of Experimental Conditions

SCR			
	Cond '1'	Cond '4'	Cond '12'
Guilty	0.36 (0.47)	0.74 (0.49)	0.85 (0.45)
Innocent	0.09 (0.32)	0.23 (0.24)	-0.05(0.19)
RLL			
	Cond '1'	Cond '4'	Cond '12'
Guilty	0.24 (0.35)	0.45 (0.39)	0.51 (0.41)
Innocent	-0.16 (0.24)	0.09 (0.19)	0.005 (0.28)
Combined Measure			
	Cond '1'	Cond '4'	Cond '12'
Guilty	0.57 (0.72)	1.13 (0.75)	1.41 (0.60)
Innocent	-0.07 (0.47)	0.33 (0.30)	-0.04 (0.35)

Table 3: Correct detection rates for "Guilty" and "Innocent" participants, computed using the Lykken's Scoring Procedure, for each physiological measure in the three experimental conditions.

SCR

	Innocent	Guilty
Single question repeated twelve times	91.66% (N=12)	50% (N=24)
Four questions repeated three times	83.3% (N=12)	75% (N=24)
Twelve different questions	100% (N=12)	79.16% (N=24)

RLL

	Innocent	Guilty
Single question repeated twelve times	100% (N=12)	33.33% (N=21)
Four questions repeated three times	100% (N=12)	40% (N=20)
Twelve different questions	100% (N=12)	80% (N=20)

SCR+RLL

	Innocent	Guilty
Single question repeated twelve times	100% (N=12)	47.61% (N=21)
Four questions repeated three times	91.66% (N=12)	80% (N=20)
Twelve different questions	100% (N=12)	95% (N=20)

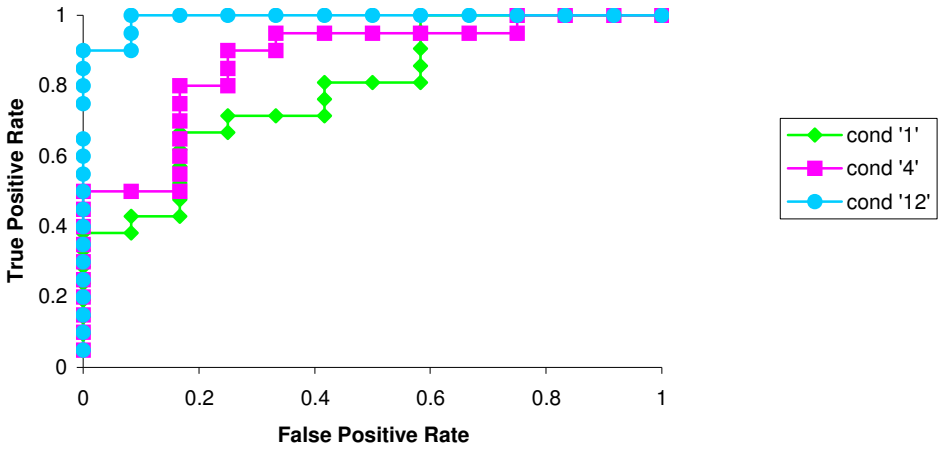
Table 4: Areas under the ROC curves and 90% CI for the areas produced by each measure as a function of Experimental conditions

COMBINED MEASURES	RLL	SCR	
0.794 (0.673 - 0.915)	0.849 (0.746 - 0.953)	0.677 (0.536 - 0.818)	Single question repeated twelve times
0.871 (0.772 - 0.969)	0.829 (0.706 - 0.952)	0.813 (0.703 - 0.922)	Four questions repeated three times
0.992 (0.974 - 1.000)	0.829 (0.721 - 0.937)	0.986 (0.968 - 1.000)	Twelve different questions

Figure Captions

Figure 1: Areas under the ROC curves computed for the combined measure under the three experimental conditions.

Figure 1


True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
cond '1'
cond '4'
cond '12'