

Final Report: ***Growth Curve Analysis of Polygraph Data***

Grant No. DASW01-03-1-0001
Department of the Army

21 May, 2003

Elizabeth Lockette
Department of Educational Psychology
University of Utah
Salt Lake City, UT 84112
voice: (801) 581-7148
fax: (801) 581-5566

John C. Kircher, Ph.D.
Department of Educational Psychology
University of Utah
Salt Lake City, UT 84112
voice: (801) 581-7130
fax: (801) 581-5566
email: kircher@gse.utah.edu

Abstract

Growth curve analysis was used in the present study to test if skin conductance responses habituate during polygraph examinations, if the responses of guilty and innocent subjects habituate at different rates, and if differential rates of habituation can be used to improve the accuracy of computer diagnoses of truth and deception. The data for the present project came from two previously conducted mock crime experiments. One study was conducted at the University of Utah with 84 participants. The other study was conducted at the FBI Academy with 120 participants. Half of the subjects in each experiment were guilty of committing a mock theft, half were innocent, and all subjects were offered a monetary bonus to convince the polygraph examiner of their innocence. Although there were significant and substantive differences between the guilty and innocent groups in rates of habituation, the resulting parameter estimates did not significantly improve the accuracy of computer decisions. Alternative models of growth for skin conductance and models of cardiovascular and respiration responses were not explored that might increase the discrimination between truthful and deceptive individuals.

Background

Probable-lie Polygraph Tests

The Probable-Lie Test (PLT) is the most common type of polygraph test for criminal investigation in the United States (Office of Technology Assessment, 1983). The PLT contains relevant, probable-lie, and neutral questions. Relevant questions pertain to the matter under investigation; e.g., “Did you rob the 7-11 on May 18th?” Probable-lie questions address a general content area that is related to the crime but excludes the particular matter under investigation; e.g., “Before the age of 23, did you ever take something that didn’t belong to you?” Neutral questions serve as buffer items; e.g., “Do you live in the United States?” All test questions are reviewed with the subject prior to the test. Relevant questions are reviewed first, and subjects generally answer the relevant questions “No.” Probable-lie questions are reviewed next, and neutral questions are reviewed last. When the probable-lie questions are introduced, the subject is led to believe that admission to those questions would raise doubts about the person’s veracity concerning the crime – that they would be viewed as the type of person who would steal something and lie about it. The manner in which probable-lie questions are introduced is designed to embarrass or intimidate the subject into answering “No.” If the subject answers “Yes” to a probable-lie question, the question is reworded slightly to elicit a “No” response from the subject; e.g., “Other than what you told me, before the age of 23, did you ever take something that didn’t belong to you?” Even if a probable-lie question is reworded, it is difficult or impossible for subjects to answer such a question truthfully

with a “No.” The PLT is so-named because the answers to probable-lie questions by all subjects are probably false. The neutral questions are reviewed last.

The PLT is based on the assumption that subjects will react most strongly to the type of question that poses the greatest perceived threat to their appearing truthful on the test (Podlesny & Raskin, 1977). Guilty subjects answer the relevant questions deceptively. Because the relevant questions pertain directly to the matter under investigation, guilty subjects are expected to react more strongly to them than to the probable-lie questions. Conversely, innocent subjects answer the relevant questions truthfully, but are likely to be deceptive or unsure about the truthfulness of their answers to the probable-lie questions. Therefore, innocent subjects are expected to react more strongly to the probable-lie questions than to the relevant questions. It is expected that guilty and innocent subjects will show their weakest reactions to the neutral questions, although reactions to the neutral questions typically are not evaluated. Table 1 contains an example question list for a PLT concerning the theft of a ring.

Table 1. Example question list for a probable-lie test

1. (Buffer) Do you understand that I will ask only the questions that we have discussed?
2. (Sacrifice Relevant) Do you intend to answer truthfully each question about the theft of the ring?
3. (Neutral) Do you live in the United States?
4. (Probable-lie) Before the age of 23, did you ever take something that didn't belong to you?
5. (Relevant) Did you take the ring from the secretary's desk?

6. (Neutral) Is your first name Richard?
7. (Probable-lie) Between the ages of 12 and 23, did you ever break a law, rule, or regulation?
8. (Relevant) Did you take that ring?
9. (Neutral) Is today Tuesday?
10. (Probable-lie) During the first 23 years of your life, did you ever lie to get out of trouble?
11. (Relevant) Do you know where the ring is now?

In the example sequence, decisions would be based on pairwise comparisons of physiological reactions to probable-lie and relevant questions at positions 4 and 5, 7 and 8, and 10 and 11. If reactions generally were stronger to the relevant than to the probable-lie questions, the subject would be called deceptive to the relevant questions. If reactions to the probable-lie questions were greater, the subject would be considered truthful to the relevant questions. If there were little difference between reactions to probable-lie and relevant questions, the test would be inconclusive.

The polygraph records subjects' respiration, electrodermal, and cardiovascular responses to test questions. Test questions are presented at a rate of one question every 25 to 30 seconds. The entire set of questions is presented several times, and each repetition of the question sequence provides a chart. If the test is inconclusive after three repetitions of the question sequence (charts), the polygraph examiner often will run one or two additional charts. Between charts, the examiner deflates the blood pressure cuff for recording cardiovascular activity and gives the subject a one to three minute break. To maintain the salience of probable-lie questions, during the break, the examiner may ask

about one of the probable-lie questions; e.g., “Did something come to mind when I asked you if you ever broke a law, rule, or regulation?” The position of each relevant question remains constant across charts, but neutral and probable-lie questions are rotated among their respective positions such that each relevant question is preceded by each neutral and each probable-lie question at least once (Raskin & Honts, 2002).

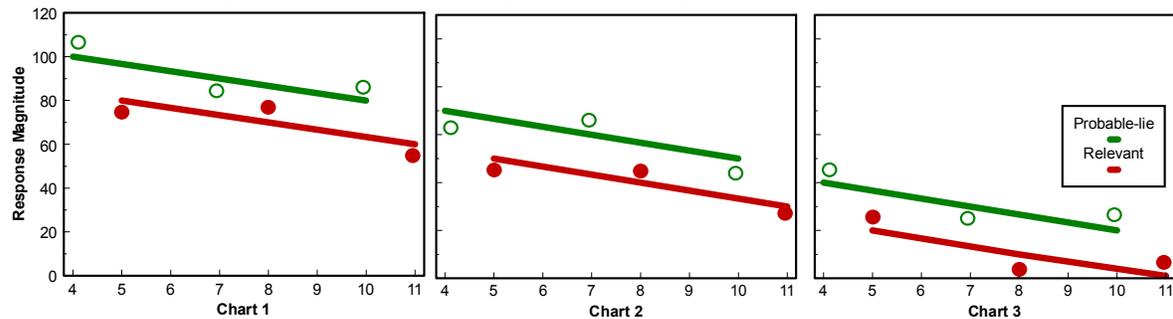
In realistic mock crime experiments, well-trained polygraph interpreters and computers reach decisions in 85% to 90% of cases, and about 90% of those decisions are correct (Raskin et al., 1999). However, polygraph decisions are based exclusively on accumulated (mean) differences in responses to probable-lie and relevant questions. No human or computer scoring technique considers the possibility that truthful and deceptive subjects show different patterns of change in response magnitude over questions within charts or across charts. If the trajectories of growth curves vary as a function of deceptive status, then they could provide a new source of diagnostic information that is at least partially independent of differences in mean levels. Estimates of slope parameters then might be used in combination with level differences to improve discrimination between truthful and deceptive subjects.

Growth Curve Analysis

A growth curve in the present study was the line of best fit to a series of observed measurements of SC amplitude. One growth curve represented the linear change in SC amplitude over the three PL questions at positions 4, 7, and 10 in the first chart. Another growth curve showed the change in SC amplitude over the three relevant questions at positions 5, 8, and 11 in the first chart. Growth curves were similarly defined for subsequent charts.

Figure 1 shows a set of six growth curves for three charts for a hypothetical subject. The circles represent observed measurements of SC amplitude, and the lines are the fitted growth curves.

Figure 1. Fitted growth curves for a hypothetical subject



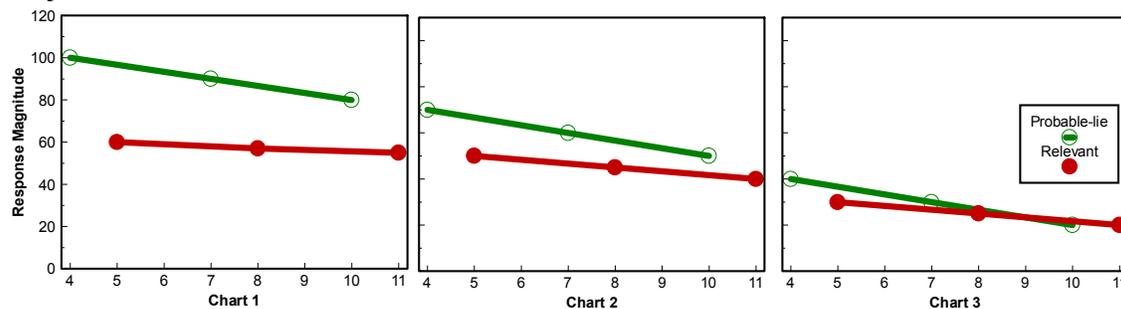
In an analysis of growth curves, the Y-intercept of each growth curve serves as a dependent variable. A subject with three charts would provide three intercepts for the PL growth curves and three intercepts for the three relevant question growth curves. To improve interpretability, increase the stability of parameter estimates, and reduce multicollinearity, the mean of X is often subtracted from each of the original scores to center the time variable. In the present example, test questions appeared at positions 4, 5, 7, 8, 10, and 11. Question position (X) would be centered about the mean position ($M = 7.5$), and the resulting values on the X-axis for a chart would be -3.5, -2.5, -.5, .5, 2.5, and 3.5, respectively. Centering puts the Y-intercept at the center (mean) of the growth curve, and makes the Y-intercept the mean level of the growth curve. Subsequent tests of Y-intercepts then become tests of the mean levels of the growth curves.

In Figure 1, the intercepts drop over Charts, and the mean intercept for PL questions is greater than the mean intercept for relevant questions. If this pattern were characteristic of most subjects in an experiment, one would expect a main effect of Charts on intercepts and a main effect of Question Type on intercepts.

An analysis of growth curves treats the slope of each growth curve as a second dependent variable. In Figure 1 above, all of the slopes are negative, and they are all equal. Since the lines are parallel, there is no change in the slope of the growth curve over Charts (no main effect of Charts on slopes) or over types of questions (no main effect of Question Type on slopes).

Another pattern of responses to PL and relevant questions is shown in Figure 2. Again, the Y-intercepts drop over charts (main effect of Charts on intercepts). There also is a mean difference between the intercepts for PL and relevant questions that favors the PL questions. Finally, there is a Chart X Question Type interaction because the difference between the intercepts for PL and relevant questions decreases over charts.

Figure 2. Fitted growth curves for probable-lie and relevant question for a hypothetical subject



The slope for PL questions is steeper than the slope for relevant questions (main effect of Question Type on slopes). Responses to PL questions habituated more rapidly than responses to relevant questions. Since the mean slope for PL and relevant questions is constant over charts, there is no main effect of Charts on slopes. Finally, although the difference between the intercepts changes over charts, the difference between the slopes does not (no Chart X Question Type interaction effect on slopes).

Comparison of Growth Curve Analysis and Repeated Measures ANOVA

Traditionally, repeated measurements of physiological reactions to probable-lie and relevant questions are analyzed with repeated measures analysis of variance (RMANOVA; e.g., Podlesny & Raskin, 1978). In fact, there is a close relationship between growth curve analysis and traditional RMANOVA. According to Bryk and Raudenbush (1992), the methods yield the same conclusions when the data are completely crossed and balanced and the RMANOVA assumptions are met.

However, there are conceptual and practical advantages in using hierarchical models to analyze growth curves. The primary purpose of the present study was to determine if slope estimates for individual subjects are diagnostic, because human and computer methods of chart analysis currently do not use them. The hierarchical linear model (HLM) provides estimates of slopes for individual subjects, whereas RMANOVA does not. RMANOVA models variation in growth as an interaction of Groups and Occasions, and parameter estimates for individual subjects are not readily available.

Second, RMANOVA requires that measurement Occasions be completely crossed with Persons. In contrast, HLM treats measurement occasions as though they were nested within persons. In the present study, probable-lie and relevant questions were nested in Charts, and Charts were nested in Subjects. The latter approach is more accommodating as it allows for unequal spacing between measurement occasions and for unequal numbers of observations across people. In the present study, every subject in an experiment provided the same number of observations, and the spacing between questions was approximately equal and constant across all subjects. However, Kircher et al. (2001) obtained five charts per subject, whereas Podlesny and Kircher (1999) obtained only four charts per subject. Although we did not do so, HLM would allow us to

combine the data sets from the two experiments into a single analysis. Such an analysis would not be possible with RMANOVA without dropping the fifth chart (20% of the data) for the subjects in one experiment.

Third, HLM integrates measurement theory and traditional hypothesis testing. HLM partitions the observed variance in intercepts or slopes into true score (reliable) variance and error variance. At each stage of model development, the analysis software reports the proportion of unexplained variance in the outcome measure that is reliable. As independent variables are added to the model to test hypotheses, the proportion of true score variance explained by the independent variable is estimated (effect size), and the proportion of residual variance in the outcome measure that is reliable is also reported. As explanatory variables are added to the hierarchical model, more and more of the reliable variance is explained. When the reliable (true score) variance approaches zero, there is no need to add any additional explanatory variables to the model, since all of the variance that can be explained (true score variance) has been explained. Although effect size statistics may be obtained following RMANOVA, investigators rarely do so. In addition, traditional effect size statistics provide no indication of whether reliable variance remains in model residuals. If so, then factors other than those included in the model affect the dependent variable. Better theory and more research would be warranted.

A RMANOVA of the data for a simple laboratory study of polygraph techniques would require four factors, and all factors but Subjects would be considered fixed. The design would contain one between-group factor (Guilt) with two levels (guilty and innocent), and three within-subject factors: Charts with three to five levels, Question

Position with three levels (QP), and Question Type (QT) with two levels (probable-lie and relevant). The linear model for this RMANOVA would include 8 random effects (error terms; Subject main and interaction effects) and 16 fixed effects. The fixed effects would include the grand mean (μ), main effects for Guilt, Chart, QP, and QT, six two-way interaction terms (Guilt*Chart, Guilt*QP, Guilt*QT, Chart*QP, Chart*QT, and QP*QT), four three-way interactions (Guilt*Chart*QP, Guilt*Chart*QT, Guilt*QP*QT, and Chart*QP*QT), and one four-way interaction (Guilt*Chart*QP*QT). The hierarchical model for this design provided a statistical test for each of the 16 fixed parameters in this linear model. In contrast to RMANOVA, a hierarchical analysis would also provide tests of the random effects.

In the present study, hierarchical linear models of growth were developed using procedures described in the text by Bryk and Raudenbush (2002) and the HLM Version 5 computer program (Bryk, Raudenbush, & Congdon, 2002). HLM provided estimates of growth parameters (intercepts and slopes) for each type of question (PL and relevant) and each chart. HLM also provided statistical tests for the following research hypotheses:

1. Physiological responses habituate within charts.
2. Physiological responses habituate across charts.
3. Habituation within charts varies linearly as a function of charts.
4. Guilt moderates the effects of Question Type on mean levels of growth curves for PL and relevant questions.
5. Guilt moderates habituation rates.
6. Within-chart habituation varies as a function of Guilt and Question Type.
7. The between-chart habituation varies as a function of Guilt and Question Type.

8. Reliable variance among individuals remains in means and slopes after controlling for Guilt, Chart, Question Type, and Question Position.

Conditional on finding differences between guilty and innocent subjects in the slopes of their growth curves, our plan was to determine if the slopes could be used to improve the accuracy of computer diagnoses of truth and deception. We planned to develop a multiple regression equation to predict Guilt (0/1) from differences in the levels of growth curves for probable-lie and relevant questions (intercepts) and then to add slope estimates to the regression equation to test if:

9. Slope parameters can be used to increase the accuracy of computer diagnoses of truth and deception.

Methods

The present project used 84 subjects from one polygraph experiment (Study A; Kircher et al., 2001) and 120 subjects from another experiment (Study B; Podlesny & Kircher, 1999). Both studies used a mock crime paradigm, and in both studies, equal numbers of male and female subjects were randomly assigned to guilty and innocent treatment conditions. All subjects were recruited from the community, paid for their participation, and offered a substantial monetary bonus to convince the polygraph examiner of their innocence. Both samples were diverse in terms of age, ethnicity, and socioeconomic status.

Three hundred and thirty-six subjects participated in Study A at the University of Utah (Kircher et al., 2001). That study investigated effects of a pretest demonstration of polygraph accuracy on subsequent detection rates. Guilty subjects received tape-recorded instructions to wait for a secretary to leave her office unattended, find a purse in

her desk, and take \$20 from a wallet in her purse. A senior graduate student or a post doctorate fellow collected five charts of physiological data from each subject.

Differences between the two examiners in Study A were assessed with a mixed model Examiner X Guilt X Sex ANOVA. Examiner was random and Guilt and Sex were fixed factors. Using an alpha of .20, the main effect of Examiner was not significant and Examiner did not interact with Guilt or Sex. Therefore, Examiner was omitted as a factor in the present study.

Only half of the 336 subjects in Study A received PLTs, and nonstandard procedures were used in two of four PLT treatment conditions that affected the accuracy of the test. Therefore, the present study included only subjects in the standard PLT control groups (n=60) and another PLT condition that varied in a minor way from the control condition and did not affect the accuracy of the test (n=24).

One hundred and twenty subjects participated in Study B at the FBI Academy in Quantico, VA (Podlesny & Kircher, 1999). Study B was designed to evaluate a new method for measuring blood pressure. Programmed guilty subjects took \$10 from a purse in a waiting room, denied having taken the money, and took a PLT from a psychophysicologist 3 to 14 days later. Physiological measures included respiration, skin conductance (SC), electrocardiogram, and either the cardiograph (n=40) or arterial finger blood pressure (BP) (n=40), or both (n=40). Four charts of physiological data were collected from each subject. Results revealed that diastolic BP was highly correlated with the current measure of cardiovascular activity, and systolic BP was marginally more diagnostic of truth and deception than the current measure of cardiovascular activity. Further tests revealed that the method of recording cardiovascular activity had no

discernable effect on the diagnostic validity of any of the other channels of recorded physiological activity.

Skin Conductance Measurements

Response Curves. From the series of digitized polygraph signals, response curves were generated for SC. The SC response curve was defined by the series of stored samples that began at question onset of each probable-lie or relevant question and ended 20 s later.

Feature Extraction. Peak amplitude was extracted from the SC response curve. To measure peak amplitude, low points in the response curve were identified as changes from negative or zero slope to positive slope, and high points in the response curve were identified as changes from positive slope to zero or negative slope. The difference between each low point and every succeeding high point was computed. Peak amplitude was defined as the greatest such difference.

Within-subject Standardization. To remove variance among individuals in basal levels of physiological activity and reactivity, the repeated measurements of SC amplitude were transformed to z-scores within each subject. For example, in Study A, there were 30 measurements of SC amplitude since there were 3 probable-lie and 3 relevant questions (6 questions) on each of 5 charts. For each subject, the mean and standard deviation of the 30 measurements were used to transform each of the 30 raw scores to 30 z-scores.

Hierarchical Linear Model

A hierarchical model with three levels provided estimates of changes in SC amplitude over Question Positions and over Charts. HLM required a different data file

for each level. The level-1, level-2, and level-3 data were organized as shown in Table 3. The level-1 data file contained as many rows as there were test questions (e.g., 6 questions/chart X 5 charts/subject X 84 subjects = 2520). The level-2 data file contained as many rows as there were charts (e.g., 5 charts/subject X 84 subjects = 420), and the level-3 contained as many rows as there were subjects in the experiment (e.g., $N = 84$).

Table 3a. Organization of level-1 data file

Subject (k)	Chart (j)		Question Type		Question Position (i)		Measure (Y_{ijk})	
		Centered	Label		Index	Centered	Raw	Z
1	1	-2	PL1	1	4	-3.5	76	1.2
1	1	-2	R1	-1	5	-2.5	42	-3
1	1	-2	PL2	1	7	-.5	81	1.4
1	1	-2	R2	-1	8	.5	33	-4
1	1	-2	PL3	1	10	2.5	69	.8
1	1	-2	R3	-1	11	3.5	41	-.1
1	2	-1	PL2	1	4	-3.5	58	.3
1	2	-1	R1	-1	5	-2.5	38	-3
1	2	-1	PL3	1	7	-.5	71	1.0
1	2	-1	R2	-1	8	.5	44	-.1
1	2	-1	PL1	1	10	2.5	53	.3
1	2	-1	R3	-1	11	3.5	29	-5
...
1	5	2	PL2	1	10	2.5
1	5	2	R3	-1	11	3.5
2	1	-2	PL1	1	4	-3.5
2	1	-2	R1	-1	5	-2.5
...

Table 3b. Organization of the level-2 data file

Subject (k)	Chart (j)	
	Index	Centered
1	1	-2
1	2	-1
1	3	0
1	4	1
1	5	2
2	1	-2
2	2	-1

2	3	0
...

Table 3c. Organization of the level-3 data file

Subject (k)	Guilt
1	1
2	-1
3	-1
4	1
...	...

Level-1 Models. At level 1, the linear model was:

$$\text{Level 1} \quad Y_{ijk} = \pi_{0jk} + \pi_{1jk} \text{QP} + \pi_{2jk} \text{QT} + \pi_{3jk} \text{QP*QT} + e_{ijk}$$

where

Y_{ijk} was a SC response for question position i , chart j , and subject k

π_{0jk} was the mean level of the growth curves for PL and relevant questions for chart j and subject k . π_{0jk} was estimated from the mean of the six measured responses on a chart, and it provided a global measure of response amplitude for a chart.

QP was a question position centered about the mean position ($M = 7.5$).

π_{1jk} was the effect of Question Position for chart j and subject k . π_{1jk} was the mean slope of the growth curves for PL and relevant questions for chart j and subject k . Conceptually, π_{1jk} provided an overall measure of habituation within a chart.

QT was a dichotomous variable that distinguishes between PL questions (coded 1) and relevant questions (coded -1).

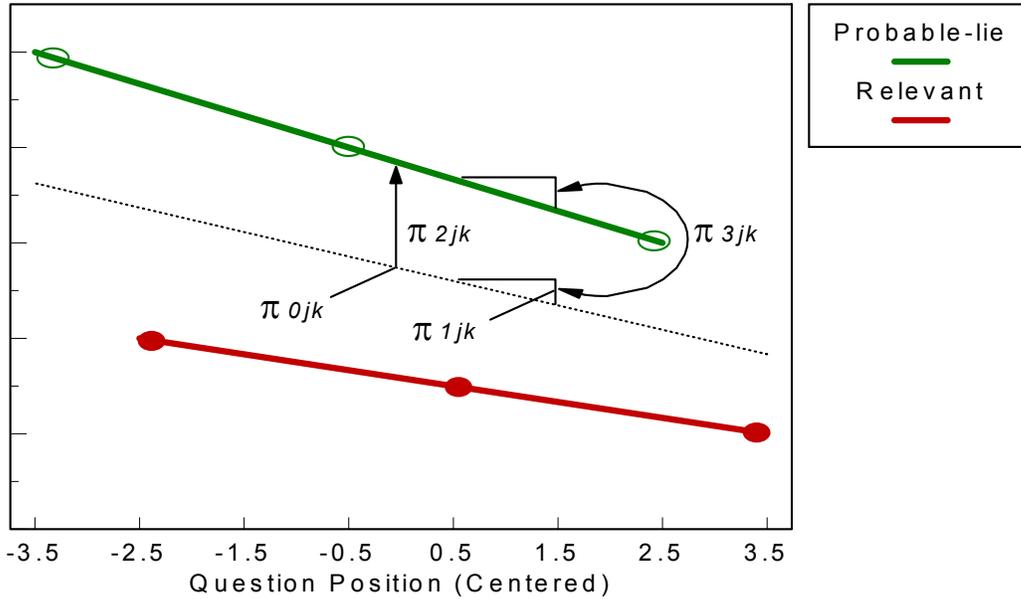
π_{2jk} Effect of Question Type. π_{2jk} was the difference between the level of the growth curve for PL questions and the mean level of the growth curves for chart j and subject k (π_{0jk}). The PLT predicts that π_{2jk} will be positive for innocent subjects and negative for guilty subjects.

QP*QT was a vector of the cross-products of QP and QT and was used to measure the interaction effect (π_{3jk}).

- π_{3jk} was the effect of Question Position X Question Type interaction. π_{3jk} was difference between the slope of the growth curve for PL questions and the mean slope of the growth curves for chart j and subject k (π_{1jk}). π_{3jk} differed from zero to the extent that the within-chart slope for one type of question (habituation rate) differed from the slope for the other type of question. Specifically, π_{3jk} for chart j and subject k was positive when responses to relevant questions habituated more rapidly than responses to PL questions, and it was negative when responses to PL questions habituated more rapidly.
- e_{ijk} was the within-chart error. e_{ijk} was the deviation of the measured SC response at position i from the fitted growth curve for chart j and subject k

Note that each effect (π) has subscripts j and k . Since subscripts appear for charts (j) and subjects (k), there were as many level-1 regression models and estimates of each effect (π) as there were charts in the experiment (e.g., 5 charts per subject X 84 subjects = 420 regression equations). Each regression equation could be used to ‘predict’ the responses to the three PL questions and the three relevant questions in a particular chart j for particular subject k . Figure 3 provides a graphical representation of model parameters that would be estimated for one hypothetical chart.

Figure 3. Effects measured by a level-1 model. The dotted line represents the mean of the growth curves for probable-lie and relevant questions.



Level-2 Models. At level-2, parameters were estimated for the following models:

$$\begin{aligned} \text{Level 2} \quad \pi_{0jk} &= \beta_{00k} + \beta_{01k} \text{CHART} + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} + \beta_{11k} \text{CHART} + r_{1jk} \\ \pi_{2jk} &= \beta_{20k} + \beta_{21k} \text{CHART} + r_{2jk} \\ \pi_{3jk} &= \beta_{30k} + \beta_{31k} \text{CHART} + r_{3jk} \end{aligned}$$

where,

- β_{00k} was the mean of all SC responses for subject k .
- β_{01k} was the linear change in the mean within-chart SC response across charts for subject k (habituation across charts).
- β_{10k} was the mean slope of within-chart growth curves for subject k (mean habituation within charts).
- β_{11k} was the linear change in the mean within-chart slope across charts for subject k .
- β_{20k} was (half) the mean difference between PL and relevant questions for subject k .
- β_{21k} was the linear change in the difference between probable-lie and relevant questions across charts for subject k .
- β_{30k} was the mean within-chart difference between the slopes of the growth curves for probable-lie and relevant questions for subject k .

β_{31k}	was the linear change in the within-chart difference between the slopes of the growth curves for probable-lie and relevant questions across charts for subject k
r_{jk}	were deviations between fitted values and observed π_{jk}

There were four sets of level-2 regression equations, one for each growth parameter in the level-1 model. The dependent variables for the level-2 models were the mean level (π_{0jk}) and the effects of QP, QT, and the QP*QT interaction in the level-1 model (π_{1jk} , π_{2jk} , and π_{3jk} , respectively). When five charts were available for person k , there were five measures of the mean SC response (π_{0jk}) for person k , one for each chart. The explanatory variable CHART in each level-2 equation was centered about the mean chart number ($M = 3$). For five charts, the values of CHART were -2, -1, 0, 1, and 2, as shown in Table 3b. The k subscript for a β indicates that the β varied over subjects; that is, there were as many regression equations for a given level-2 outcome measure as there were subjects in an experiment.

Level-2 Model for π_{0jk} . π_{0jk} was the mean of all of the SC responses to PL and relevant questions within a chart. HLM fit a line to the five values of π_{0jk} for person k . The slope of the line for subject k was β_{01k} and the intercept was β_{00k} . Since CHART was centered, β_{00k} was the mean of all measured responses for person k .

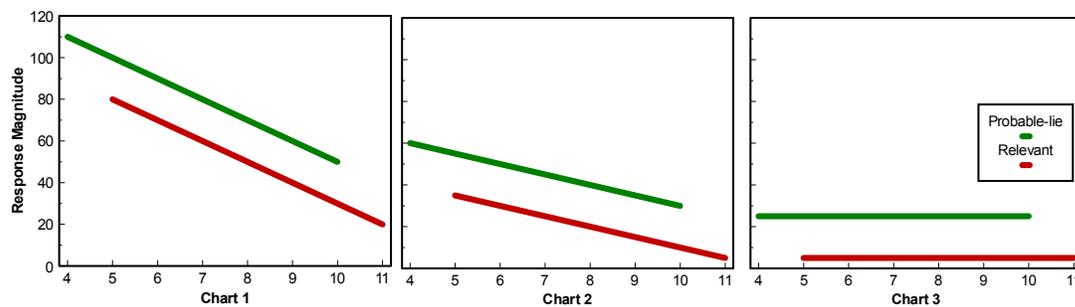
Habituation across charts was indicated by a negative value of β_{01k} . Figure 1 shows one possible pattern of habituation of SC responses between charts. In Figure 1, the mean response, π_{0jk} , decreased over charts 1, 2, and 3. The decrease in π_{0jk} over charts would be indicated by a negative value of β_{01k} .

Level-2 Model for π_{1jk} . A second level-2 equation was specified for the mean within-chart slope (π_{1jk} ; see dotted line in Figure 3). HLM fit a line to the five estimates

of π_{1jk} for person k . The intercept of that line was β_{10k} , and the change in the slopes over charts was indicated by β_{11k} . Since CHART was centered, the intercept, β_{10k} , was the mean of all within-chart slopes for subject k . In Figure 1, all within-chart slopes were negative and they were equal. Therefore, a line connecting the within-chart slopes over charts would be flat ($\beta_{11k} = 0$) and the level of that line would be negative ($\beta_{10k} < 0$).

Figure 4 shows a different pattern of habituation over charts. In Figure 4, the mean within-chart habituation, π_{1jk} , gets progressively less negative over charts. By Chart 3, the mean within-chart slope has increased to zero. In this case, the mean within-chart slope would be negative ($\beta_{10k} < 0$), and the change in within-chart slopes would be positive ($\beta_{11k} > 0$).

Figure 4. A pattern of habituation that shows a progressive increase in the within-chart slopes over charts



Level-2 Model for π_{2jk} . π_{2jk} reflected the within-chart difference between the responses to PL and relevant questions for subject k (see Figure 3). β_{20k} in the level-2 model for π_{2jk} was the mean value of π_{2jk} for subject k , and β_{21k} was the linear change in π_{2jk} over charts. In Figure 2, the mean difference between PL and relevant questions was positive ($\beta_{20k} > 0$) for this subject, despite the lack of any appreciable difference in Chart

3. The PLT predicts that β_{20k} will be positive for innocent subjects and negative for guilty subjects.

In Figure 2, the difference between PL and relevant questions decreases over charts. Therefore, the slope of a line fit to the differences would be negative ($\beta_{21k} < 0$). If this pattern were characteristic of innocent subjects, then it would be easier to verify a person's truthfulness on the first chart than the third.

Level-2 Model for π_{3jk} . π_{3jk} was a measure of the (linear X linear) interaction between Question Position and Question Type. π_{3jk} would be zero if the growth curves for PL and relevant questions within a chart were parallel, as shown in Figure 1; π_{3jk} would be negative if responses habituate more rapidly to PL questions, as shown in Figure 2; and π_{3jk} would be positive if responses habituate more rapidly to relevant questions.

The level-2 model for π_{3jk} provides the mean QP*QT interaction across the charts for subject k , β_{30k} . The model for π_{3jk} also provides the change in π_{3jk} over charts, β_{31k} . Essentially, β_{31k} reflects the three-way interaction between Charts, Question Position, and Question Type for subject k . In the parlance of HLM, β_{31k} is a cross-level interaction effect because a level-2 factor (CHART) moderates a level-1 effect. β_{11k} and β_{21k} also would be considered measures of cross-level interaction.

In Figure 2, responses to PL questions always habituate more quickly than do responses to relevant questions. Thus, the subject mean value of π_{3jk} would be negative ($\beta_{30k} < 0$). However, the difference between the slopes for PL and relevant questions is constant over charts. Therefore, $\beta_{31k} = 0$.

Residuals for Level-2 Models. The residuals (r_{jk}) for the level-2 models are deviations between the estimated π_{jk} and the value predicted by the level-2 regression model. The within-subject variance among the observed residuals for a level-2 model may be pooled across subjects and tested for statistical significance. A significant result would indicate that the level-2 model, which includes only a linear effect of Charts, does not account for all the reliable within-subject variance among charts in the associated growth parameter. Such a finding might indicate the presence of quadratic or higher-order trend components.

Level-3 Models. The level-3 models were as follows:

$$\begin{aligned} \text{Level 3} \quad \beta_{00k} &= \gamma_{000} + \gamma_{001} \text{ GUILT} + u_{00k} \\ \beta_{01k} &= \gamma_{010} + \gamma_{011} \text{ GUILT} + u_{01k} \\ \beta_{10k} &= \gamma_{100} + \gamma_{101} \text{ GUILT} + u_{10k} \\ \beta_{11k} &= \gamma_{110} + \gamma_{111} \text{ GUILT} + u_{11k} \\ \beta_{20k} &= \gamma_{200} + \gamma_{201} \text{ GUILT} + u_{20k} \\ \beta_{21k} &= \gamma_{210} + \gamma_{211} \text{ GUILT} + u_{21k} \\ \beta_{30k} &= \gamma_{300} + \gamma_{301} \text{ GUILT} + u_{30k} \\ \beta_{31k} &= \gamma_{310} + \gamma_{311} \text{ GUILT} + u_{31k} \end{aligned}$$

where,

γ_{000}	was the grand mean response amplitude
γ_{001}	was the main effect of Guilt
γ_{010}	was the main effect of Chart
γ_{011}	was the Chart X Guilt interaction
γ_{100}	was the main effect of Question Position
γ_{101}	was the Question Position X Guilt interaction
γ_{110}	was the Question Position X Chart interaction
γ_{111}	was the Question Position X Chart X Guilt interaction
γ_{200}	was the main effect of Question Type
γ_{201}	was the Question Type X Guilt interaction
γ_{210}	was the Question Type X Chart interaction
γ_{211}	was the Question Type X Chart X Guilt interaction

γ_{300}	was the Question Position X Question Type interaction
γ_{301}	was the Question Position X Question Type X Guilt interaction
γ_{310}	was the Question Position X Question Type X Chart interaction
γ_{311}	was the Question Type X Question Position X Chart X Guilt interaction
$u_{..k}$	were the deviations between fitted values and the obtained $\beta_{..k}$

At level 3, each level-2 effect served as a dependent variable. GUILT was a dichotomous variable that distinguished between innocent (coded 1) and guilty subjects (coded -1). Consequently, the intercept in each level-3 model ($\gamma_{..0}$) was the grand mean of $\beta_{..k}$ across all subjects. The $u_{..k}$ were the deviations of subjects' $\beta_{..k}$ about their respective group means. Significant within-group variance of estimated $u_{..k}$ would suggest that other characteristics of subjects such as age or sex might be added to the level-3 model to explain the variance among subjects within the two treatment conditions.

Proportions of Reliable Variance Explained

The HLM program reports maximum likelihood estimates of true-score variance as well as the ratio of true-score variance to observed-score variance for each outcome measure (reliability). Ordinarily, a hierarchical analysis begins with the analysis of an unconditioned or null model with no independent variables in the level-2 or level-3 equations. Analysis of the unconditioned model provides baseline measures of reliability as well as statistical tests to determine if the variance within or between subjects is significant. If the variance of a growth curve parameter is not significant, then there is no need to develop a model with independent variables to explain that variance. It is only when there are reliable differences among measurement units that explanatory variables are added to the regression equation to account for those differences.

If an independent variable is added to a level-2 or level-3 regression equation and its coefficient is significant, the proportion of variance explained by the independent variable may be assessed by comparing the variances of model residuals before and after the independent variable has been included in the model. In the present study, the proportion of reliable within-subject variance explained by CHART was assessed as follows:

$$\frac{\text{VAR}(r_{jk}) \text{unconditioned} - \text{VAR}(r_{jk}) \text{conditioned}}{\text{VAR}(r_{jk}) \text{unconditioned}}$$

where VAR (r_{jk}) unconditioned was the estimated reliable variance among model residuals without CHART in the level-2 equation, and VAR (r_{jk}) conditioned was the estimated reliable variance among model residuals with CHART in the level-2 equation. Likewise, the proportion of reliable between-subject variance explained by GUILT was assessed as follows:

$$\frac{\text{VAR}(u_{jk}) \text{unconditioned} - \text{VAR}(u_{jk}) \text{conditioned}}{\text{VAR}(u_{jk}) \text{unconditioned}}$$

where VAR (u_{jk}) unconditioned was the estimated reliable variance among subjects about the grand mean without GUILT in the level-3 equation, and VAR (u_{jk}) conditioned was the estimated reliable variance among subjects about their respective treatment group means.

Results

The analysis of SC data was conducted in two phases. In the first phase, an unconditioned model was developed and the model was simplified. In the second phase, independent variables were added to the level-2 and level-3 equations to answer our

research questions. The analyses were conducted separately for Study A and Study B to assess the consistency of findings across experiments.

Phase I

An unconditioned model was analyzed with no level-2 or level-3 explanatory variables to determine if there was reliable within-subject variance among charts ($\text{VAR}(r_{.jk})$) or among subjects ($\text{VAR}(u_{.k})$). The unconditioned model was as follows:

$$\text{Level 1} \quad Y_{ijk} = \pi_{0jk} + \pi_{1jk} (\text{QP}) + \pi_{2jk} (\text{QT}) + \pi_{3jk} (\text{QP*QT}) + e_{ijk}$$

$$\begin{aligned} \text{Level 2} \quad \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} + r_{1jk} \\ \pi_{2jk} &= \beta_{20k} + r_{2jk} \\ \pi_{3jk} &= \beta_{30k} + r_{3jk} \end{aligned}$$

$$\begin{aligned} \text{Level 3} \quad \beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{10k} &= \gamma_{100} + u_{10k} \\ \beta_{20k} &= \gamma_{200} + u_{20k} \\ \beta_{30k} &= \gamma_{300} + u_{30k} \end{aligned}$$

β_{00k} was the mean level of the growth curves for subject k . For example, in Study A, there were 5 charts and there were growth curves for probable-lie and relevant questions for each chart. In that case, β_{00k} was the mean level of the 10 growth curves for subject k . Since the repeated measures for each subject had been transformed to z-scores, and the mean of any set of z-scores is zero, the observed estimate of β_{00k} was exactly zero for every subject. Therefore, the results from each study indicated that the grand mean level (γ_{000}) did not differ from zero and there was no reliable variance among individuals in their values β_{00k} . Since all β_{00k} were zero, the mean, γ_{000} , was zero, there were no deviations about the mean (all u_{00k} were zero), and the $\text{VAR}(u_{00k})$ was zero (see the level-3 equation for β_{00k}). Consequently, β_{00k} was dropped from the model.

π_{3jk} was half the difference between the slopes of the growth curves for probable-lie and relevant questions for chart j and subject k (see Figure 2). The mean of the four or five π_{3jk} for subject k was β_{30k} . To determine if π_{3jk} would remain in the level-1 model three tests were conducted. The first test was to determine if there was reliable variance among the π_{3jk} within subjects. A χ^2 test indicated that the variance of π_{3jk} about the subject mean (β_{30k}) or $\text{VAR}(r_{3jk})$, was not significant. Thus, the difference between the slopes of the growth curves for probable-lie and relevant questions did not vary over charts.

Next, a χ^2 test was conducted to determine if there were reliable differences among subjects in their mean values of π_{3jk} . The variance of β_{30k} about the grand mean (γ_{300}) was $\text{VAR}(u_{30k})$, and the χ^2 test of $\text{VAR}(u_{30k})$ was not significant. Therefore, differences among subjects in values of β_{30k} were not significant. Finally, γ_{300} was tested and the grand mean QP*QT interaction effect did not differ from zero. Since the grand mean did not differ from zero and there was no reliable variance in π_{3jk} within or between subjects, the decision was made to drop QP*QT from the level-1 model. The same analyses, results, and conclusions regarding the QP*QT interaction were obtained for Study A and Study B. Since QP and QT were centered and balanced, QP*QT was orthogonal to QP and to QT, and the presence or absence of QP*QT in the level-1 model had no effect on the parameter estimates for QP or QT.

Within-Subject Variances

χ^2 tests were conducted to test if there was reliable within-subject variance among the levels and slopes of growth curves for the four or five charts. The $r_{.jk}$ were the deviations of π_{0jk} , π_{1jk} , and π_{2jk} about their respective subject means β_{00k} , β_{10k} , and β_{20k} .

The variances of the r_{jk} were significant in both Study A and Study B. Table 4 presents the results of the χ^2 tests for r_{0jk} , r_{1jk} , and r_{2jk} . These findings indicate that the within-chart level of the growth curves (π_{0jk}), the mean within-chart slope of the growth curves (π_{1jk}), and the difference between the levels of growth curves for PL and relevant questions (π_{2jk}) changed over charts.

Table 4. χ^2 tests of within-subject variances

Parameter	Study	Variance	Reliability	df	χ^2	p-value
r_0	A	.174	.615	336	1069	.00
	B	.153	.559	360	1085	.00
r_1	A	.004	.186	336	507	.00
	B	.008	.220	360	606	.00
r_2	A	.072	.398	336	622	.00
	B	.029	.195	360	547	.00

Between-Subject Variance

χ^2 tests were also conducted to test for reliable between-subject variances. Table 5 summarizes the results. β_{10k} was the mean within-chart slope of the growth curves for subject k . A χ^2 test indicated that the variance of β_{10k} about the grand mean γ_{100} , $\text{VAR}(u_{10k})$, was not significant for Study A or Study B. Since there was no reliable variance among subjects in mean within-chart slopes of growth curves, it would not be possible to use β_{10k} to distinguish between guilty and innocent subjects.

β_{20k} was half the mean difference between the levels of growth curves for PL and relevant questions for subject k . The PLT predicts that innocent subjects will show stronger reactions to PL than to relevant questions, and guilty subjects will show stronger reactions to the relevant questions. Therefore, positive values of β_{20k} were expected for innocent subjects, negative values were expected for guilty subjects, and substantial

variance in β_{20k} was expected. As predicted, the χ^2 test of the variance of β_{20k} about its grand mean γ_{200} , $\text{VAR}(u_{20k})$, was significant for Study A and Study B.

Table 5. χ^2 tests of between-subjects variances

Parameter	Study	Variance	Reliability	df	χ^2	p-value
u_{10k}	A	.0010	.191	83	90	0.29
	B	.0012	.131	119	131	0.21
u_{20k}	A	.0402	.527	83	175	0.00
	B	.0828	.689	119	386	0.00

Phase II

The hierarchical model was revised based on the results obtained in Phase I. The QP*QT factor was removed from the level-1 model, and β_{00k} was removed from the level-2 model for π_{0jk} . CHART was added to each level-2 model because the Phase I χ^2 tests for $r_{.k}$ were significant. In addition, Guilt was added to the level-3 models to provide tests of the research hypotheses.

$$\text{Level 1} \quad Y_{ijk} = \pi_{0jk} + \pi_{1jk} (\text{QP}) + \pi_{2jk} (\text{QT}) + e_{ijk}$$

$$\begin{aligned} \text{Level 2} \quad \pi_{0jk} &= \beta_{01k} (\text{CHART}_{jk}) + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} + \beta_{11k} (\text{CHART}_{jk}) + r_{1jk} \\ \pi_{2jk} &= \beta_{20k} + \beta_{21k} (\text{CHART}_{jk}) + r_{2jk} \end{aligned}$$

$$\begin{aligned} \text{Level 3} \quad \beta_{01k} &= \gamma_{010} + \gamma_{011} (\text{GUILT}_k) + u_{01k} \\ \beta_{10k} &= \gamma_{100} + \gamma_{101} (\text{GUILT}_k) + u_{10k} \\ \beta_{11k} &= \gamma_{110} + \gamma_{111} (\text{GUILT}_k) + u_{11k} \\ \beta_{20k} &= \gamma_{200} + \gamma_{201} (\text{GUILT}_k) + u_{20k} \\ \beta_{21k} &= \gamma_{210} + \gamma_{211} (\text{GUILT}_k) + u_{21k} \end{aligned}$$

Table 6 summarizes the results of analyses of the simplified hierarchical model that address the first seven research questions. The “Yes” or “No” answer to each research question is based on the outcome of a two-tailed t-test of the associated parameter at $p < .05$. Where possible, for significant effects, Table 6 reports the

proportion of total variance that was true-score variance before CHART was added to the level-2 model or before GUILT was added to the level-3 model (reliability). The last column reports the proportion of that true-score variance that was explained by a factor or cross-level interaction.

1. *Do physiological responses, Y_{ijk} , habituate within charts?*

π_{1jk} was the mean slope of the growth curves for probable-lie and relevant questions for chart j and subject k (see Figure 3). The mean within-chart slope for subject k was β_{10k} , and the grand mean within-chart slope was γ_{100} . Examination of the results in Table 5 revealed that the estimate of γ_{100} was significant for Study A, $t(83) = -2.13$, $p < .05$, and Study B, $t(119) = -4.69$, $p < .01$, and it was negative. SC responses habituated within-charts. However, the effects were small. The proportion of observed score variance explained by Question Position was only .02 in Study A and .04 in Study B.

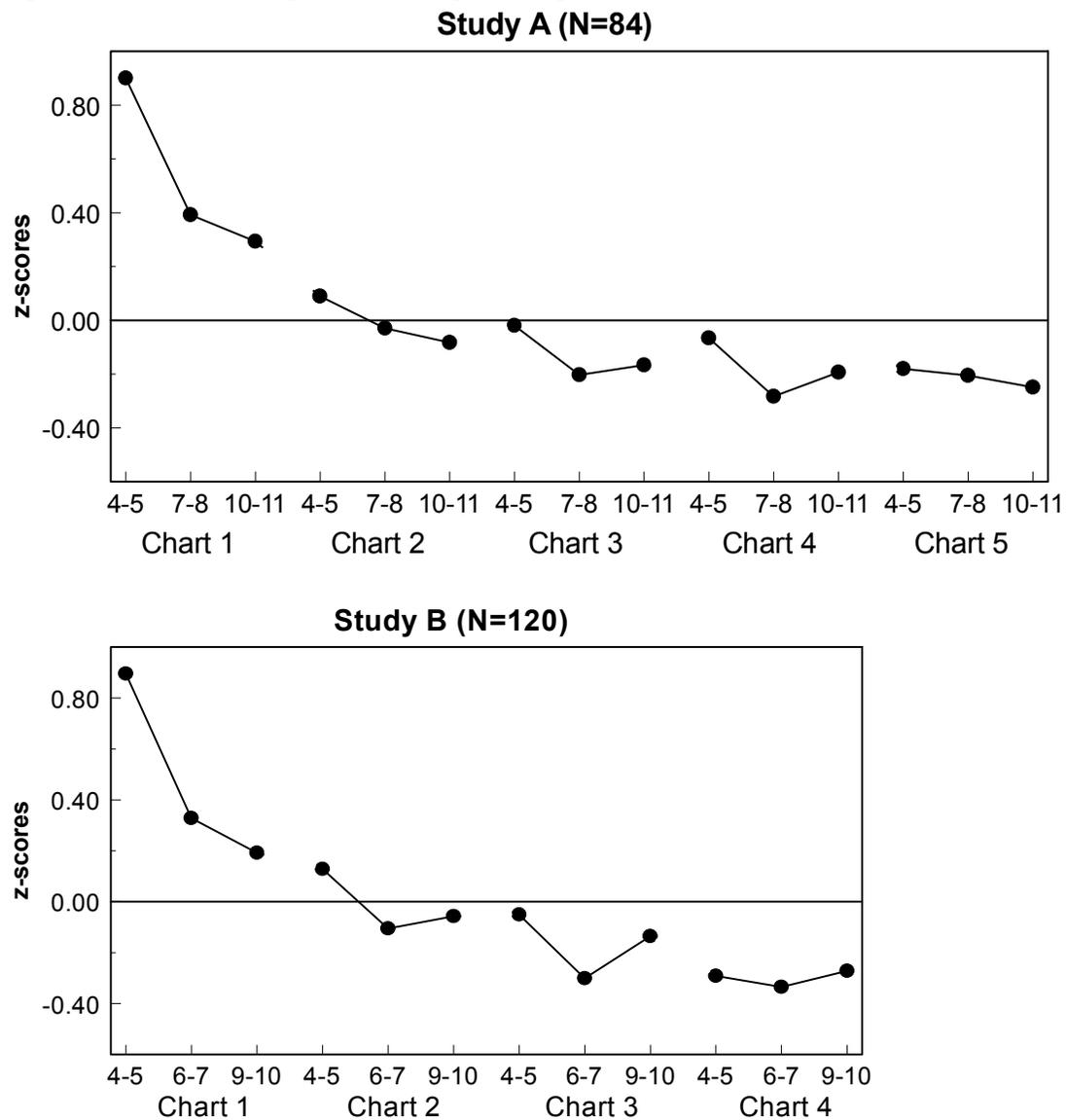
Figure 5 shows the mean z-score for each question position across the five charts in Study A as well as the mean z-scores across the four charts in Study B. The data in Figure 5 reveal a systematic decline in the amplitude of SC responses within the first two charts. Thereafter, the slopes of the growth curves approach zero.

Table 6. Summary of results of statistical tests of research hypotheses¹.

	Research Question	Parameter	ANOVA effect	Study	Answer	Estimate	Proportion true score variance	Proportion true score variance explained
1	Do physiological responses habituate within charts?	γ_{100}	QP	A	Yes	-0.017		
				B	Yes	-0.042		
2	Do physiological responses habituate across charts?	γ_{010}	Chart	A	Yes	-0.118	0.615	0.085
				B	Yes	-0.280	0.559	0.830
3	Does within-chart habituation vary over charts?	γ_{110}	QP X Chart	A	Yes	0.015	0.186	0.892
				B	Yes	0.056	0.220	0.760
4	Does Guilt moderate the effects of Question Type on mean levels of growth curves?	γ_{201}	Guilt X QT	A	Yes	0.374	0.527	0.766
				B	Yes	0.400	0.689	0.456
5a	Does Guilt moderate within-chart habituation rates?	γ_{101}	QP X Guilt	A	No	-0.012	-	-
				B	No	0.021	-	-
5b	Does Guilt moderate changes in within-chart habituation rates over charts?	γ_{111}	QP X Chart X Guilt	A	No	-0.008	-	-
				B	No	-0.030	-	-
5c	Does Guilt affect the rate of habituation over charts?	γ_{011}	Chart X Guilt	A	No	-0.068	-	-
				B	No	0.060	-	-
6	Do within-chart habituation rates vary as a function of Guilt and Question Type?	γ_{301}	QP X QT X Guilt	A	No	-	-	-
				B	No	-	-	-
7	Does Guilt affect between-chart habituation rates to PL and relevant questions?	γ_{211}	QT X Chart X Guilt	A	Yes	-0.127	0.398	0.365
				B	Yes	-0.104	0.195	0.304

¹Note: Only linear effects were considered for factors with more than 1 degree of freedom (QP and Charts).

Figure 5: Mean SC amplitude over question positions



2. Do physiological responses, Y_{ijk} , habituate across charts?

π_{0jk} was mean level of the growth curves for chart j and subject k (see Figure 3). The linear change in π_{0jk} from one chart to the next for subject k was β_{01k} , and the grand mean change in the level of the growth curves from one chart to the next for all subjects was γ_{010} .

Examination of the results for Question 2 in Table 6 reveals that there was a significant drop in the level of the growth curves over charts in Study A, $t(83) = -5.31$, $p < .05$, and in Study B, $t(119) = -8.44$, $p < .05$. In Study A, SC amplitude dropped .12 standard deviations between charts, and in Study B, SC amplitude dropped .28 standard deviations between charts. The proportions of reliable variance in π_{0jk} in the two studies were comparable, but CHARTS accounted for considerably more of the reliable variance in Study B (.83) than in Study A (.08). A straight line better fit the four data points (charts) in Study B than the five points in Study A. Examination of Figure 5 suggests that there was a strong quadratic component to the growth curve defined by the five chart means.

3. *Does within-chart habituation vary over charts?*

The data in Figure 5 indicate that within-chart slopes varied systematically across charts. Specifically, habituation in the first chart was quite dramatic, and there was progressively less evidence of habituation in latter charts.

π_{1jk} was the mean within-chart slope for chart j and subject k (see Figure 2). The linear effect of CHART on within-chart slopes for subject k was β_{11k} , and the grand mean effect of CHART on within-chart slopes across all subjects was γ_{110} .

The results in Table 6 indicate that the within-chart slope varied linearly as a function of charts. The slope changed positively at a mean rate of .02 standard deviations in Study A, $t(83) = 2.51$, $p < .05$, and at a rate of .06 standard deviations in Study B, $t(119) = 5.06$, $p < .05$. Since γ_{110} was positive, it indicated that the within-chart slope became less negative and approached zero over the course of the polygraph examination.

Although relatively little of the observed variance in π_{1jk} was reliable, CHARTS explained most of the reliable variance.

4. Does Guilt moderate the effects of Question Type on mean levels of growth curves?

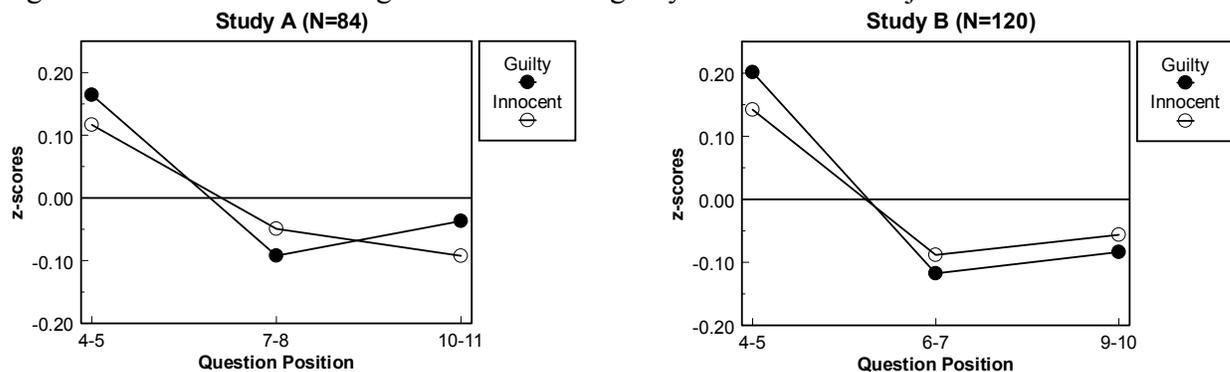
π_{2jk} was half the difference between the level of the growth curves for probable-lie and relevant questions for chart j and subject k (see Figure 3). The mean effect of Question Type across charts for subject k was β_{20k} .

Decisions concerning deception on a polygraph test are currently based on mean differences in physiological responses to probable-lie and relevant questions; i.e., β_{20k} . As expected, the effect of Guilt on the difference between probable-lie and relevant questions (γ_{201}) was significant in Study A, $t(83) = 8.46$, $p < .05$, and in Study B, $t(119) = 7.72$, $p < .05$.

5a. Does Guilt moderate within-chart habituation rates?

Figure 6 displays pooled within-chart growth curves for guilty and innocent subjects in Study A and Study B. Habituation was evident within charts for guilty and innocent subjects, but there was little difference between guilty and innocent subjects in the rate of habituation. These observations were confirmed by statistical analysis.

Figure 6. Mean within-chart growth curves for guilty and innocent subjects



π_{1jk} was the mean within-chart slope for chart j and subject k , β_{10k} was the subject mean within-chart slope, and γ_{101} was effect of Guilt on those subject means. As shown in Table 5, the test of γ_{101} was not significant for Study A or Study B. There was no evidence that Guilt moderated linear growth rates within a chart.

5b. Does Guilt moderate changes in within-chart habituation rates over charts?

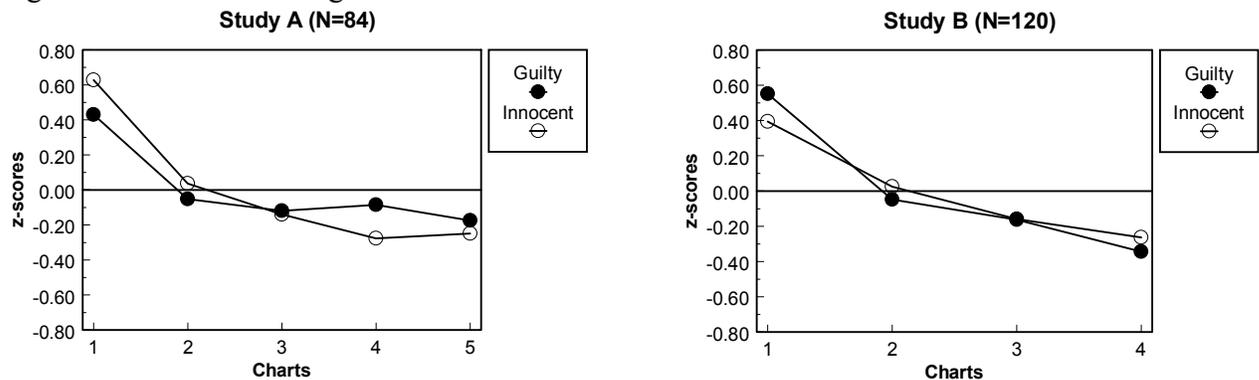
We also evaluated the possibility that guilty and innocent subjects could be distinguished in terms of the rate of change in within-chart slopes over charts. γ_{111} provided a test of the Question Position X Chart X Guilt interaction. The results in Table 6 indicate that γ_{111} did not differ from zero. There was no evidence that Guilt moderates changes in within-chart growth rates over charts.

5c. Does Guilt affect the rate of habituation over charts?

π_{0jk} was the mean level of the growth curves for chart j and subject k . β_{01k} was the slope of a line fit to the four or five values of π_{0jk} for subject k . β_{01k} provided an index of between-chart habituation, and γ_{011} provided a test of the difference between guilty and innocent subjects in their values of β_{01k} .

Figure 7 displays the mean level of growth curves over charts for Study A and Study B. Habituation between charts was evident for guilty and innocent subjects in both studies. However, the test of γ_{011} revealed no difference in the rate of habituation for guilty and innocent subjects in either study. These results suggest that Guilt does not affect habituation across charts.

Figure 7. Mean levels of growth curves over charts



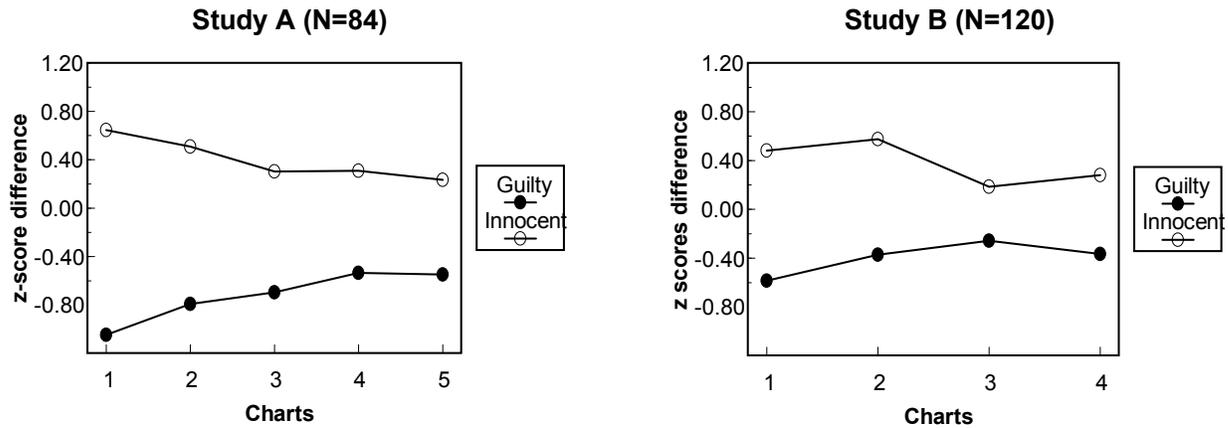
6. *Do within-chart habituation rates vary as a function of Guilt and Question Type?*

Analysis of the unconditioned model in Phase 1 indicated that the within-subject and between-subject variances associated with the Question Position X Question Type interaction ($\text{VAR}(r_{3jk})$ and $\text{VAR}(u_{30k})$) were not significant. Since there was no reliable variance in the measures of Question Position X Question Type interaction, there was no reason to test if within-chart habituation rates varied as a function of Guilt and Question Type.

7. *Does Guilt affect between-chart habituation rates to probable-lie and relevant questions?*

Figure 8 plots the *difference* between the levels of the growth curves for probable-lie and relevant questions for guilty and innocent subjects. Examination of Figure 8 indicates that the absolute difference between probable-lie and relevant questions decreased over charts.

Figure 8. Mean difference between SC responses to probable-lie and relevant questions for guilty and innocent subjects



In the hierarchical model, π_{2jk} was half the difference between the level of the growth curves for probable-lie and relevant questions for chart j and subject k (see Figure 3). The linear effect of CHART on π_{2jk} for subject k was β_{21k} . The data in Figure 8 suggest that the values of β_{21k} tended to be negative for innocent subjects and positive for guilty subjects. The test of the Question Type X Chart X Guilt interaction was a test of the difference between the mean values of β_{21k} for guilty and innocent subjects. The difference between the means was γ_{211} .

The results in Table 5 indicate that γ_{211} differed significantly from zero and was negative for Study A, $t(83) = -4.11$, $p < .05$, and in Study B, $t(119) = -3.23$, $p < .05$. The value of γ_{211} was negative because innocent subjects had high scores on GUILT (1) and negative scores on β_{21k} , whereas guilty subjects had low scores on GUILT (-1) and positive scores on β_{21k} .

8. Does reliable variance among individuals remain in means and slopes after controlling for Guilt, Chart, Question Type and Question Position?

Table 6 shows the results of χ^2 tests of the residual variances for each of the growth curve parameters in the hierarchical model. For all but two parameters, the χ^2 test indicated that the residual variance exceeded chance levels of variability. After controlling for Guilt, Chart, Question Type, and Question Position, reliable variance remained in most of the growth parameters that might be explained by other variables in future studies.

Table 6. χ^2 tests of residual variances for growth curve parameters in the hierarchical model

Parameter	Study	Residual Variance	df	χ^2	p-value	Does Variance Remain?
r_0	A	0.159	336	891.19	0.00	Yes
	B	0.026	360	561.70	0.00	Yes
r_1	A	0.000	252	407.82	0.00	Yes
	B	0.002	240	485.09	0.00	Yes
r_2	A	0.046	252	564.60	0.00	Yes
	B	0.020	240	526.27	0.00	Yes
u_{01}	A	0.014	82	129.00	0.00	Yes
	B	0.038	118	282.94	0.00	Yes
u_{10}	A	0.001	82	105.70	0.04	Yes
	B	0.003	118	155.38	0.01	Yes
u_{11}	A	0.001	82	134.95	0.00	Yes
	B	0.002	118	139.54	0.09	No
u_{20}	A	0.009	82	106.12	0.04	Yes
	B	0.045	118	271.93	0.00	Yes
u_{21}	A	0.004	82	105.00	0.04	Yes
	B	0.003	118	116.30	> 0.50	No

9. *Can slope parameters can be used to increase the accuracy of computer diagnoses of truth and deception.*

Aside from mean differences between probable-lie and relevant questions (β_{20k}), the only slope parameter that reliably distinguished between guilty and innocent subjects was β_{21k} . β_{21k} was the linear change in the difference between SC responses to probable-lie and relevant questions over charts (see Figure 7). For each study, a traditional hierarchical regression analysis was performed to test if this growth parameter could be

used in combination with other physiological measures to improve discrimination between the groups.

The statistical model currently used by the Computerized Polygraph System (CPS) to discriminate between truthful and deceptive subjects uses mean differences in the magnitude of respiration, SC, and cardiovascular responses to probable-lie and relevant questions (Kircher & Raskin, 2001). Those three measures were extracted from the polygraph charts and were used as predictor variables in a multiple regression equation to predict a dichotomous variable that distinguished between guilty (coded -1) and innocent (coded 1) subjects.

Ordinary least squares estimates of β_{21k} were then added to the regression equation, and the regression coefficient for β_{21k} was tested for statistical significance.

The results are summarized in Table 7.

Table 7. Point-biserial (r_{pb}) and standardized regression coefficients for traditional physiological measures and a growth parameter (β_{21k})

Physiological Measure	Study A		Study B	
	r_{pb}	Std Regression Coefficient	r_{pb}	Std Regression Coefficient
SC amplitude	.76**	.67**	.57**	.30**
BP amplitude	.40**	.05	.45**	.11
Respiration	.23*	.14*	.53**	.29**
est β_{21k}	-.41**	-.14	-.30**	-.12

** $p < .01$

* $p < .05$

As expected, bivariate correlations with Guilt were significant for all traditional measures of mean differences in the magnitude of physiological responses to probable-lie and relevant questions. In addition, the change in the difference between SC responses to probable-lie and relevant questions from one chart to the next (β_{21}) was significantly

correlated with the criterion in both studies. However, the unique contribution of β_{21} to a regression equation that contained the traditional measures did not achieve statistical significance for either study. Even if the standardized regression coefficients for β_{21} had been significant, the addition of β_{21} increased the R^2 from .60 to .62 in Study A and from .39 to .41 in Study B. Correlations among the physiological measures and the criterion are presented in Appendix A.

Discussion

The results of the present study confirmed several predictions, the most important of which was that truthful and deceptive individuals react differently to probable-lie and relevant questions. In two independent samples, we found that innocent subjects reacted more strongly to probable-lie comparison questions, and guilty subjects reacted more strongly to relevant questions. The present study also demonstrated that SC responses habituated over the course of a polygraph examination. Habituation was evident within and between charts. It demonstrated that SC responses of guilty and innocent subjects to probable-lie and relevant questions habituated at different rates. Responses to probable-lie questions habituated faster for innocent subjects, and responses to relevant questions habituated faster for guilty subject. Consequently, differences between SC responses to probable-lie and relevant questions became smaller and approached zero over charts. Since differences between probable-lie and relevant questions decreased over charts, SC data collected near the beginning of the polygraph test may be more diagnostic than those collected near the end. Finally, growth curve analysis revealed diagnostic differences in the rates of habituation. However, when used in combination with traditional measures of

mean differences in responses to probable-lie and relevant questions, habituation rates did not significantly improve the accuracy of computer classifications.

Growth curve analysis produced results that were consistent with already established techniques for assessing truth and deception. One research question asked if the mean levels of growth curves were affected by the Question Type X Guilt interaction. Considerable prior research predicted effects of Guilt on within-subject differences between probable-lie and relevant questions. Indeed, polygraph decisions are based on such differences. Although that finding was not new, the HLM analysis did provide useful psychometric information about differential reactivity that is not commonly assessed. For example, Guilt accounted for 73% of the reliable variance in differential reactivity in Study A and 45% of the reliable variance in Study B. Thus, there was reliable variance in differential reactivity that was not due entirely to the subjects' deceptive status. Other individual differences, such as sex, age, intelligence, or interactions of such factors with Guilt, affected subjects' differential reactivity to probable-lie and relevant questions. Knowledge of major source(s) of variance in differential reactivity other than Guilt could be used to develop and test theory and to improve the accuracy of diagnoses. For example, further study might reveal that differential electrodermal reactivity to probable-lie and relevant questions is more diagnostic for young males with low to moderate intelligence than for older exceptionally bright females.

The finding that Guilt accounted for less of the reliable variance in Study B (45%) than Study A (73%) might be due to differences in subject characteristics or aspects of the research design. For example, Study B contained a higher percentage of Blacks

participants (20%) than did Study A (2%). In Study B, guilty subjects committed the mock crime and returned three days to two weeks later for their polygraph examination. In Study A, subjects reported immediately for their polygraph examination.

In the present study, growth curves pooled across probable-lie and relevant questions did not distinguish between guilty and innocent subjects. Since there was no evidence that simple habituation rates could be used to distinguish between the groups, there appears to be no advantage in retaining separate growth curves for probable-lie and relevant questions. Mean differences between probable-lie and relevant questions and changes in differences across charts appear to capture all of the diagnostic variance in SC measures. As such, the hierarchical model could be simplified by using difference scores as the dependent variable and dropping Question Type as a factor from the level-1 model.

Over the course of the present study, other interesting questions arose that could be addressed with growth curve analysis. For example, polygraph examiners sometimes refer to probable-lie questions between charts to focus attention on them and reduce the risk of false positive outcomes; e.g., “Did anything come to mind when I asked if you ever lied to get out of serious trouble?” (Raskin & Honts, 2001). Although responses to probable-lie questions habituate within charts, they may recover (dishabituate) somewhat between charts. Piecewise growth curve analysis may be used to test the hypothesis that such statements by the examiner produce a discontinuity in the habituation trajectory for probable-lie questions between charts (Bryk & Raudenbush, 1991; J. Butner, personal communication, July 2002).

A lack of discontinuity between charts might argue for further simplification of the model. Charts could be omitted as a factor in the model. Growth curves would then

be defined by repeated measures from the first presentation of a probable-lie or relevant question on the first chart to the last presentation on the last chart. By omitting Charts as a factor, the analysis would proceed as a two-level hierarchical model rather than a three-level model. A quadratic growth parameter then might be added to the level-1 model.

It is worthwhile to reiterate that it would be possible to combine the data from Study A and Study B and perform a single analysis. Study would be added as a between group factor at level 3, and it would allow for tests of main and interaction effects of Study on the dependent variable. Such an analysis would be impossible with repeated measures ANOVA because different numbers of charts were obtained from the subjects in the two experiments, and RMANOVA requires that measurement occasions be crossed with subjects. In HLM, measurement occasions are nested within subjects rather than crossed with subjects. Thus, if the number of measurement occasions varies over subjects, it is a matter of dealing with unequal n's, not missing values.

If habituation reduces the effectiveness of the PLT, then efforts may be made to retard its effects. For example, the wording of questions may be modified slightly between charts. In this way, subjects would have to process the meaning of each new question before they answer. Even if the wording of only one or two questions were changed, it would require subjects to pay more attention to all of the questions and may reduce the effects of habituation.

It is important to note that the present findings might not generalize to polygraph examinations conducted on actual criminal suspects. The present study was conducted using data from two mock crime experiments. Although there was consistency in the findings from the two experiments, the findings might differ if growth curve analyses are

conducted with data from actual criminal suspects. In addition, our growth curve analyses were limited to SC measurements. It is unknown if the pattern of habituation observed for SC responses would be found for other physiological measures.

In conclusion, the results of growth curve analysis revealed that in laboratory experiments, SC responses habituate over the course of probable-lie polygraph test. Although differential rates of habituation were diagnostic, when combined with traditional measures of mean differential reactivity, they did not improve the accuracy of computer decisions.

Growth curve analysis may be used to test a number of interesting hypotheses that were not evaluated in the present study. For example, it might be used to determine if changes in physiological measures other than SC can be used to improve the accuracy of computer decisions. It might be used to test if the adverse effects of habituation can be reduced by making small changes in the wording of questions over charts and increasing the cognitive demands of the task. Alternatively, it might be used to test if statements made by the polygraph examiner to enhance the signal value of probable-lie questions before each chart function as expected and interrupt the trajectory of the growth curves for probable-lie questions. Moreover, if such statements affect only innocent subjects, measurements of those effects would be diagnostic and might add to a computer model for detecting deception.

Appendix A

Intercorrelations among physiological measures and the guilt/innocence criterion for Study A (above the principal diagonal; N = 84) and Study B (below the principal diagonal; N = 120)

	Guilt	SC	Cardiograph	Respiration	Est β_{2IK}
Guilt	1.00	.76	.40	-.23	-.41
SC	.57	1.00	.49	-.11	-.37
Cardiograph	.45	.60	1.00	.02	-.17
Respiration	-.53	-.55	-.44	1.00	-.09
Est β_{2IK}	-.30	-.33	-.27	.18	1.00

Correlations above the principal diagonal beyond +/- 0.21 were significant at $p < .05$

Correlations below the principal diagonal beyond +/- 0.18 were significant at $p < .05$

References

- Bain, L. J., Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*, Second Edition. Boston: PWS-Kent Publishing Company.
- Bryk, A. S. & Raudenbush, S. W., (1992). *Hierarchical Linear Models, Applications and Data Analysis Methods*. Thousand Oaks, California: Sage Publications, Inc.
- Hays, W.L. (1994). *Statistics*, 5th ed. Orlando, Florida: Harcourt College Publications.
- Kircher, J. C., Packard, R. E., Bell, B. G. & Bernhardt, P. C. (2001). *Effects of prior demonstrations of polygraph accuracy on outcomes of probable-lie and directed-lie polygraph tests (Grant No. DoDPI97-P-0016)*. Final report to the U. S. Department of Defense. Salt Lake City: University of Utah, Department of Educational Psychology.
- Office of Technology Assessment (1983). *Scientific validity of polygraph testing: A research review and evaluation: A technical memorandum*. OTA-TM-H-15, NTIS order #PB84-181411. Washington, DC: U.S. Government Printing Office.
- Podlesny, J. A. & Kircher, J. C. (1999). The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations: Experimental comparison with the cardiograph method. *Forensic Science Communication*, 1(3), 1-17.
- Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 15, 344-359.
- Raskin, D. C., Honts, C. R., Amato, S., & Kircher, J. C. (1999). The case for the admissibility of the results of polygraph examinations. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *The scientific evidence manual (Volume 1) 1999 Pocket Part*. St. Paul: West Publishing Co.
- Raskin, D. C. & Honts, C. R. (2001). The comparison-question polygraph technique. In M. Kleiner (Ed.). *Handbook of Polygraph Testing*. London: Academic Press.
- Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models, Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, California: Sage Publications, Inc.