Admissability of Polygraph Tests:

The Application of Scientific Standards Post-*Daubert*

Leonard Saxe

Brandeis University

Waltham, MA

Gershon Ben-Shakhar

The Hebrew University of Jerusalem

Jerusalem, Israel

Abstract

The U.S. Supreme Court's decision in *Daubert* modernized the long-standing *Frye* precedent and requires courts to make scientific judgments. Courts, however, are not well-equipped to parse scientific arguments and the *Daubert* criteria offer only a rudimentary framework for decision-making. To illustrate the problems, as well as possible ways for courts to deal with scientific evidence, the paper focuses on the controversy over admissibility of polygraph (so-called "lie detector") test evidence. Application of the Daubert criteria for assessing whether polygraph test results can stand as admissible evidence are considered. The concepts of "reliability" and "validity", as used in the behavioral sciences, are discussed in relation to polygraph testing and the key question suggested by *Daubert* as to whether extant research actually tests the accuracy of polygraphy is examined. This discussion demonstrates the difficulties in attempting to apply the *Daubert* criteria, because validity is a very broad concept with both theoretical and empirical aspects, and because proper empirical tests of any scientifically-based technique must satisfy complex methodological criteria. The present analysis demonstrates that although the validity of polygraph test results has been examined across many studies, none of them satisfies the necessary criteria, and therefore, accuracy rates of polygraph test results are unavailable. If *Daubert* criteria are to be applied, social scientists and courts need to develop a common language. Although it is unreasonable to expect judges to develop the skills of expert scientists, they must become educated science consumers. There is some evidence, at least in the case of polygraph testing, that courts are making these complex judgments and that justice is being served.

Admissability of Polygraph Tests:

The Application of Scientific Standards Post-*Daubert*

For nearly 75 years, psychologically-based evaluations of deceptiveness -- so-called lie-detector tests, now referred to as polygraphs -- have been at the forefront of legal controversy about the admissibility of scientific evidence. The 1923 *Frye* precedent (*Frye v. United States)*, which was for many years the standard for admissibility of scientific evidence, arose from a question about the admissibility of an early version of present-day polygraph tests. The Frye rule conditioned legal admissibility on acceptance of a technique by the relevant scientific community and it set the precedent until relatively recent changes in the rules of evidence. In 1993, when the Supreme Court of the United States decided *Daubert*, the limited scope of *Frye* was expanded in several critical ways. This expanded role for courts in making scientific judgments raises a host of questions about how theoretical and empirical research can be distilled for legal decision-making. This is particularly so for behavioral science evidence – it is not surprising that an issue about the accuracy of a psychological test was grist for the long-standing precedent – and a decade of progress in the study of human behavior has made the issue even more complicated.

The papers in the present *Special Issue* on the *Daubert* decision (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993) share a common focus on the dilemma faced by courts in applying a broad set of criteria for determining the admissibility of scientific evidence. As Chief Justice Rehnquist noted in his cautionary opinion in *Daubert*, “I defer to no one in my confidence in federal judges ... But I do not think

[we should impose on them] ... the obligation or the authority to become amateur scientists". Although some disagree that *Daubert* stretches courts' abilities (e.g., Blanck & Berven, 1998), in critical areas of law, Justice Rehnquist's concern has become reality. Courts are struggling to make judgements about complex validity arguments (see, e.g., Faigman, Kaye, Saks & Sanders, 1997) and, as Kraus and Sales (this volume) have noted, *Daubert* may place an insurmountable burden on courts to make decisions that they are ill-equipped to render.

The present paper focuses on the debate over the admissibility of polygraph test evidence. The introduction of polygraph test evidence has been repeatedly litigated It is an important exemplar of the type of scientific issues that that courts must address. The goal of the present paper is to parse the specific scientific questions that affect the legal assessment of "reliability" of polygraph evidence and, in part, to bridge the gap between the languages of law and science. This paper tries to demonstrate what would be required to evaluate properly a psychological technology that has potentially profound effects on the judicial process.

The present analysis, like that of Krauss and Sales (this volume) who have examined the application of the *Daubert* criteria to evidence concerning child custody, adopts an inherently skeptical position about the capacity of courts to make complex scientific judgments. In the case of the child custody determinations, there is substantial research, some of which is high quality research. But the question remains as to whether this research is applicable to the specific case to which one wishes to apply it and how to weigh the likelihood of error. Many of the same issues arise in the case of polygraph testing, although in somewhat exaggerated form. There is a question as to whether any of the extant research is directly applicable and questions as to whether it is possible to construct an error rate.The essential scientific question is

how arguments should be presented to those who must make use of scientific knowledge and how courts can be helped to understand complex research issues. The difficulty of these issues is illustrated by the decades of legal controversy over the admissibility of polygraph evidence (cf. *United States. v. Scheffer*, 1998). Although courts, almost universally, have rejected polygraph evidence, it continues to be litigated. The underlying question – essential to how one applies the *Daubert* standards -- is how should courts consider conflicting views, particularly when scholars have not fully resolved the issue. In the case of polygraph testing, even a cursory review of the literature suggests that there is intense disagreement about fundamental issues of the test's validity (e.g., contrast Iacono & Lykken, 1998 with Raskin, Honts & Kircher, 1998). Perhaps reflecting Chief Justice Rehnquist's comments about the limits to judges' scientific abilities, the criteria promulgated in *Daubert* and scientific views of validity are imprecisely matched. The resulting question is what can and should scientists tell courts as they wrestle with difficult decisions about the validity of particular psychological tests and interventions?

In the case of the polygraph tests, the Supreme Court recently ruled (*United States v. Scheffer, 1998)* that a defendant's constitutional rights were not infringed upon when a military court refused to admit polygraph results. Underlying *Scheffer*, as well as other legal considerations of polygraph evidence, is the question of whether polygraph tests are "reliable". As Justice Thomas noted in the decision for the majority, "there is simply no consensus that polygraph evidence is reliable". Justice Kennedy, in a concurring opinion, noted that "The continuing, good-faith disagreement among experts and courts on the subject of polygraph reliability counsels against our invalidating a per se exclusion …". The *Scheffer* decision will likely dampen attempts to introduce polygraph evidence (primarily by criminal

defendants), but until the validity issue is definitively resolved, judicial resources will continue to be devoted to the dispute over polygraph evidence. The goal of the present discussion is to aid both psychologists and legal professionals to understand the scientific issues that should be considered in evaluating the admissibility of evidence such as polygraph test results. The focus is on the fit between the *Daubert* standards and scientific considerations of reliability and validity.

Admissibility of Scientific Evidence

Frye, for more than 60 years the key precedent on the admissibility of novel scientific evidence, involved a procedure developed and conducted by a psychologist to assess a defendant's truthfulness (see Marston, 1917). Frye was a 19-year-old defendant charged with robbery and murder. Prior to trial, psychologist William Marston administered a procedure that he called a *systolic blood pressure deception test.* On the basis of the test, he determined that Frye was truthful when he denied involvement in the robbery and murder. The trial judge, however, refused to permit Dr. Marston to testify about the examination or to re-examine Frye in court using the test.

The appeals court affirmed the trial court's decision. They reasoned that the systolic blood pressure deception test was validated only by "experimental" (i.e., not well established) evidence and was not based on a "well-recognized scientific principle or discovery." The decision noted that, "…the underlying theory seems to be that 'truth is spontaneous, and comes without effort, while the utterance of a falsehood requires conscious effort, which is reflected in blood pressure. Although the court was able to deduce a theory, they concluded that the theory did not seem to have standing among the "psychological physiological" communities of science. The court noted: "The things from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. … Just

when a scientific principle crosses the line between experimental and demonstrable is difficult to define." It remains difficult and controversy over the validity of polygraph examinations, and their acceptance in the scientific community, has continued since *Frye* (cf. Iacono & Lykken, 1997).

The *Frye ruling* influenced how virtually all U.S. Courts have treated scientific evidence (*Daubert v. Merrell Dow, 1993*). Since 1975*,* however, the legal foundation for the introduction of any scientific evidence has been based on the Federal Rules of Evidence (FRE), in particular Rule 702. Although these rules formally apply only to federal courts, they have been widely adopted by states. Rule 702 holds that "If scientific, technical or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise." The rule, in essence, requires that the evidence be relevant and, as well, aid the jury. It is not explicitly stated, but the rule also appears to require scientific validity. An opinion, even by a qualified expert, based on invalid data would clearly not be helpful to the trier of fact.

The 1993 *Daubert* decision makes the validity issue explicit and changes the way in which federal courts must consider the scientific basis for an expert's opinion. As Justice Blackmun noted, the Rules of Evidence had "moved beyond" *Frye*. *Daubert* articulates four considerations that judges should apply in determining whether to admit expert testimony based on scientific evidence: (1) testability (or falsifiability), (2) error rate, (3) peer review and publication, and (4) general acceptance. What these factors mean, and by what standards they should be judged, have been the subject of a growing debate and set of interpretations (see, e.g., Berger, 1994; Federal Judiciary Center, 1994; Faigman et al., 1997). Although the criteria

reflect the Justices' rather sophisticated understanding of science, these factors are not easily translatable to all types of evidence; this is particularly so in the case of scientific evidence from psychology and other behavioral sciences. As is detailed below, the concepts of reliability and validity take on somewhat different meanings and, unlike the physical sciences, the criterion measures may be subjective and difficult to capture. *Daubert* expands courts' gatekeeping function and implicitly requires that they make complex scientific judgements across all fields of scientific endeavor.

The courts gate-keeping function is further complicated because additional provisions of the FRE make otherwise valid testimony improper for discussion at trial. Thus, for example, Federal Rule 403 requires that a balancing test be applied: the probative value of scientific testimony be balanced against its potential prejudicial effect. Since in a typical situation, a defendant seeks to introduce an exculpatory polygraph test, its potential to be prejudicial turns on whether the trier of fact accepts the scientific basis of the test. Thus, the evidence speaks to the ultimate issue of guilt or innocence and the jury is being asked to resolve a scientific conundrum that scientific experts have not been able to resolve (cf. Brown v. Darcy, 1986; U.S. v. Cordoba, 1998).

In *Scheffer*, the recently decided matter that concerned the use of polygraph evidence in a military court martial, the court focused on Military Rule of Evidence 707. The issue was whether the President, in fulfilling his responsibility to promul-gate rules of conduct for military trials, was "reasonable" to exclude, *per se*, the results of polygraph tests. The court, in a near-unanimous decision, ruled that the exclusion of polygraph evidence was reasonable and did not violate the defendant's right to present evidence. The decision turns on the fact that scientists do not agree

that polygraph test results are reliable, and according to a large number of experts (see, e.g., Bashore & Rapp, 1993; Ben-Shakhar & Furedy, 1990; Iacono & Lykken, 1997; Saxe, 1991b, 1994), polygraph tests have unacceptable levels of reliability and validity. This is particularly so for the typical application of polygraph testing – the Control Question Test (CQT) – that has been the focus of court reviews.

There is, however, in the voluminous literature on the validity of CQT-polygraph tests a strongly held minority position on the issue of whether such tests are valid to assess truthfulness or deception (see, in particular, Honts & Quick, 1995; Raskin, Honts, & Kircher, 1997). The debate among scientists who argue for the validity of polygraph testing and those, like ourselves, who consider CQTs unscientific has become increasingly vitriolic and polarized. When, as happened several dozen times, opponents and proponents testify against one another in a *Daubert* hearing, the court is faced with having to decide which scientific judgment to believe (see, e.g., *United States v. Cordoba*, 1998; *Commonwealth of Massachusssetts v. Woodward*, 1997).

The present review does not recap the debate that takes place in these confrontations; rather, the concern is with the principles courts should consider when assessing the merit of scientific evidence such as polygraph testing. The *Daubert* criteria provide only a rudimentary framework for the analysis of scientific evidence, particularly from behavior science research. It is important to consider how the criteria can be made operational so as to evaluate psychological tests and interventions. The present discussion is designed to bridge legal and scientific thinking which, despite the increasing use of science in courtrooms (Faigman et al., 1997), represent different modes of understanding. The law is typically built by analysis of cases, while science progresses by development of generalizations and the formulation of

theory. Communicating the social scientists' theoretical-empirical way of thinking is critical if *Daubert* is to function as an effective screen for useful evidence.

Validity and Reliability

Different modes of thought notwithstanding, science and law use different language. Legal terminology, for example, does not distinguish between the key scientific concepts of reliability and validity. Reliability and validity are complex, multidimensional concepts and encompass the methodological requirements for research that can assess scientifically based techniques. Assessing reliability and validity is essential in applying the *Daubert* criteria. The underlying question is what constitutes a proper evaluation of a technique, and the theory from which it is derived. It is not sufficient to demonstrate that a technique has been tested – the key question is whether the test is adequate and what generalizations are appropriate. As suggested by Justice Blackmun, an astronomer may have valid data about the phases of the moon. But the astronomer will not be justified in generalizing from these findings to predicting human behavior.

The psychometric and testing literature (see, e.g., Lord & Novick, 1968; Messick, 1995) refers to reliability and validity in terms of generalizability. Reliability deals with replicability (or reproducibility) of the test's results; that is, whether the results are generalizable across testing situations (e.g., when the same individual is tested several times under similar circumstances). Validity is concerned with a more conceptual or theoretical generalization, namely, the extent to which test results and their interpretation reflect the concept that was the focus of the measurement procedure. Both reliability and validity must be considered in the evaluation of any test, although validity is clearly the more important consideration.

In the case of polygraph tests, estimating their psychometric reliability would require several independent administrations of a polygraph examination with the same individuals, examined on the same issue. The reliability estimate would reflect the degree to which these independent examinations produced similar outcomes. Validity, in contrast, deals with the degree to which the outcome of the polygraph test is related to truth and deception (the putative constructs measured by the polygraph). Validity, of course, is more difficult to estimate, and depends, in part, on the precise meaning of the measured constructs which are being assessed (truth telling and deception). In turn, the meaning of these concepts is dependent on the availability of a theory (Cronbach & Meehl, 1955). Indeed, reliability is an important and necessary condition for admissibility: If two polygraph examiners investigating the same suspects, reach different conclusions, these outcomes are clearly useless for the trier of fact. Nevertheless, reliability does not guarantee that a given test is valid and different polygraphers can, in principle, produce identical *incorrect* outcomes.

Reliability of Scientific Evidence

In classical reliability theory (see, e.g., Crocker & Algina, 1986; Lord & Novick, 1968), reliability is estimated by a correlation between two sets of equivalent measurements (such as the same test administered twice or two equivalent forms of the same test). In some cases, reliability is estimated by correlating two sets of scores obtained from independent observers (or judges) who have evaluated the performance of a given group under specified conditions. Reliability estimates focus on different sources of inconsistencies, or measurement errors, and the choice of an appropriate reliability coefficient depends on the purposes of the specific measurement and on the desired range of generalizability. Sometimes, more than one type of reliability estimate is required. Indeed, in modern psychometric literature, classical reliability

theory has been replaced by generalizability theory (e.g., Brennan, 1992) which requires several estimates (generalizability coefficients) that focus on the various sources of measurement error.

Various sources of measurement error affect the interpretation of a polygraph examination and each of these needs to be assessed. Although some attempts have been made to assess the reliability of the outcomes of CQTs, none of them is sufficient. In the case of polygraph results, "test scores" are typically numerical values that reflect the differences in magnitude of physiological responding to the relevant and control questions (e.g., the "quantified" method suggested by Backster, 1963, or the more objective quantification proposed by Kircher & Raskin, 1988). These scores can, as well, be a qualitative classification of the subjects into specified categories (e.g., "deceptive", "non-deceptive" and "inconclusive"). The reliability of polygraph-based scores, whether expressed by numbers or by qualitative categories, refers to the degree to which these scores tend to be stable across measurement situations. Several methods are used to estimate stability, but two approaches are common: (a) Test the same individual twice on the same issue, using the same polygraph method with two examiners who administer the test independently; (b) Test subjects once, but have at least two independent examiners score their charts.

The use of independent experts yields reliability estimates of very limited use to evaluate psycho-physiological detection. Independent examiners could, in principle, reach a complete agreement despite a very low test-retest consistency. Such an outcome is, in fact, likely if the same polygraph school trained the examiners and if they used a quantified scoring method. Such a reliability estimate is analogous to an attempt to estimate the reliability of a multiple-choice aptitude test (e.g., the SAT) by computing the agreement between two independent scorers (which will be close to

perfect, if not perfect). This approach relates to just one source of measurement error: errors in chart scoring and interpretation.

The critical question is not whether two polygraph examiners score polygraph charts consistently, but whether the procedure as a whole -- including the construction of proper relevant and control questions -- is reliable. To design a proper generalizability study of CQT polygraph examinations, a representative sample of actual suspects undergoing criminal investigation should take several, independent polygraph examinations, such that each examiner has no information, either about the outcomes, or about the type of questions used in the other examinations. In addition several independent experts should score each chart. This design would allow for an estimation of several generalizability coefficients, each sensitive to a different source of inconsistency (i.e., measurement error): (a) between different pairs of relevant and control questions within each individual examination (equivalent to an internal consistency measure in classical reliability theory); (b) between different examinations (this would be equivalent to a "test-retest" reliability estimate in classical reliability theory; (c) in chart scoring and interpretation (equivalent to an "interjudge" reliability estimate, in classical reliability theory). Only when all of these sources of measurement error are uncovered, could one have confidence in the reliability estimate.

Unfortunately, reliability studies of polygraph-based classifications are scarce, and those that have been conducted have used only the between-examiners agreement approach (e.g., Barland, 1975; Horvath & Reid, 1971). Thus, it is impossible to conclude from available data whether, or to what extent, a given subject interrogated twice by independent examiners will be similarly classified. Moreover, in practice, it is doubtful whether it would be possible to administer several, independent CQT tests to the same individuals. Repeated testing would affect the examinees and create

completely different psychological conditions that may alter, for example, the placebo value of the test (see Saxe, 1991a). Thus, the second and third repeated CQT tests are not equivalent to the first test, and therefore it is an inappropriate method for assessing reliability.

The problems in establishing reliability are important, because according to the *Daubert* criteria (see also FRE Rule 702), a key question is whether the proffered technique can and has been tested. A number of expert witnesses regularly testify (and cite relevant studies) that the reliability of the polygraph has been tested (cf. Committee of Concerned Scientists, 1997; Honts & Quick, 1995). Yet, these studies have not been designed to estimate the most important sources of measurement error threatening the reliability of a polygraph test's results, and therefore they are inappropriate tests of polygraph's reliability. Thus, an expert may provide scientific evidence to a court, yet the evidence may not be useful for the purpose for which it is being offered, even if it has appeared in a scientific journal.

The underlying question is whether courts are able to apply the scientific and technical knowledge necessary to assess the sources of measurement error, and the type of generalizability study required to approximate them. The polygraph example demonstrates the complexity of this issue. Even social scientists, including psychophysiologists, have failed to grasp this complexity, as they regard CQT polygraphy as a reliable, though not necessarily valid test.

Validity

It may be that polygraph examiners, because of the common training received by many, can conduct and score tests similarly. But as discussed above, such reliability alone is insufficient as a criterion to evaluate the test. The central question concerns validity and the degree to which inferences made on the basis of the test

scores are accurate. Traditional approaches to validity (e.g., Cronbach & Meehl, 1955) identify several discrete types of validity, while more contemporary views (see, in particular, Messick, 1989, 1995) treat validity as a unified concept. Nevertheless, even Messick, who promotes construct validity as a unified way of viewing validity, distinguishes among several aspects. Messick defines validity as an overall evaluative judgment of the degree to which empirical evidence and theory support the adequacy and appropriateness of the interpretations and actions based on test scores. Messick's definition makes clear that the focus of validity is on the interpretations of the test scores and the actions (i.e., decisions) made on the basis of these scores. It also requires that these interpretations be supported both by theoretical rationale and by empirical data. Thus, empirical demonstrations are insufficient when no rationale can be formulated and justified. Similarly, the most convincing rationale would be insufficient without empirical evidence to support it. The present discussion considers only two aspects of validity: The external component which incorporates the traditional concept of predictive (or criterion) validity, and the substantive aspect, which refers to the theoretical rationales for the test results and their interpretations (construct validity, in the traditional approach).

Predictive validity is useful when the test in question is designed to predict future behavior. The most well known examples come from the area of personnel selection, where psychological tests are used to predict future success of potential employees in a specific job. If it can be demonstrated that test results are good predictors of occupational success, the use of this test to select candidates for a job can be justified. But predictive validity has a wider usage, because prediction does not necessarily refer to future behavior, and can be applied whenever the goal of the measurement procedure can be defined and measured independent of the test.

In the case of polygraph tests, if the goal is to classify a group of individuals into categories of “deceptive” vs. “truth-tellers”, in order to estimate the predictive or criterion validity, an independent measure of veracity (“criterion”) is needed. Validity is estimated by correlating the test’s results with the criterion measure. This type of validity assessment can be translated, in many cases, into accuracy (or error) rates, one of the *Daubert* factors.

Construct validity is more complex and refers to conceptual or theoretical generalizations. An assessment of construct validity of polygraph examinations, thus requires a theory of deception. Psychologists take different positions on fundamental issues, such as whether (or to what extent) deception is a personality trait (i.e., deceptive behavior is a stable tendency of certain individuals) or whether it is primarily determined by the situation (Saxe, 1991a). In addition, the basic nature of deception is unclear and individuals may behave in ways that might seem deceptive (e.g., make untruthful statements) without being aware (Ford, 1996). Should we label such behavior deceptive and how should we deal with self-deception? The concept "deception" may be a relative concept, and what may appear to be a complete truth to one person, can appear deceptive to another.

Since autonomic reactivity is not the behavior that one wants to measure, a theoretical link needs to be developed between what is measured by a polygraph test and the behavior one wants to predict (e.g., deception). The theory must tie deceptive behavior to psycho-physiological response patterns. Then, research is required to examine both the theory and the relationships between the outcomes of the polygraph investigation and deceptive behavior defined and measured by this theory. Unfortunately, despite long-standing interest in detection of deception, no theory establishing the relationship between physiological changes and deception exists. Although poly-

graph proponents (e.g., Raskin et al., 1997) claim that relevant questions elicit a stronger reaction than control questions for deceptive subjects, such conjecture is not a theory. As Katkin (1987) has noted, interpreting such physiological activity as deception is a judgment, not a valid interpretation of test results.

In addition to the absence of evidence of a unique physiological reaction to deception, it is clear that a subject's arousal may be affected by factors other than deception; most importantly, fear of detection. There is substantial evidence about other factors which cause arousal. For example, an extensive body of research has dealt with the relationship between stress and anxiety and changes in physiological measures, controlled by the autonomic nervous system (see, e.g., Selye, 1976). These are the measures monitored in a polygraph examination.

There is also a substantial literature associating the physiological measures monitored by the polygraph with concepts, such as novelty and surprise (e.g., Berlyne, 1960). The reaction to novelty has often been studied by psycho-physiologists in terms of the "Orienting Response" (OR). The OR describes a complex, non-specific response pattern (which includes both behavioral and physiological changes) that can be observed in humans and animals when a novel stimulus is introduced (see Kimmel, van Olst, & Orlebeke, 1979; Sokolov, 1966). Interestingly, in the context of a CQT polygraph examination, the rare, novel questions are the control rather than the relevant questions. The subject expects to be asked about the crime, and often has well-rehearsed responses to the relevant questions.

Thus, a theory of psycho-physiological detection of deception would have to explain how one can separate "deceptive responses" from the effects of stress associated with a polygraph interview, novelty effects and a host of other factors that may affect the physiological responses measured during the polygraph test. What is clear,

and is not a matter of dispute, is that the physiological changes monitored during a polygraph examination -- and which constitute the basis for the polygrapher's conclusion -- are sensitive to a host of factors, unrelated to deception (see, e.g., Katkin, 1987). Thus, for example, if during a polygraph examination examinees are presented with a surprising question, or if they hear a noise from outside the examination room, they will display the physiological reactions that are interpreted routinely by polygraph examiners as an indication of deception. Even if examinees recall or think of some exciting or frightening event from their past, they might show the same response pattern (see, e.g., Ben-Shakhar & Dolev, 1996).

What these examples reveal, to use Messick's (1995) terminology (see, also, Campbell & Fiske, 1959), is that polygraph test results lack several components of construct validity. Polygraph tests not only fail to meet tests of substantive validity (theoretical), but also lacks what Messick calls external validity and what others have referred to as convergent and discriminate validity. No demonstrations establish an empirical relationship between CQT polygraph test's outcomes and other measures of the same construct (convergent validity). Moreover, there are numerous demonstrations of the lack of discriminant validity, and the physiological changes monitored during a polygraph examination have been related to other concepts.

Thus, the two major sources of invalidity noted by Messick (1995) affect CQT polygraph testing. First, as argued above, the construct of deception is underrepresented by the polygraph test results, because there exist neither theoretical rationale, nor empirical evidence supporting the relationships between the physiological measures monitored during the CQT examination and deceptive behavior. The second major threat to validity suggested by Messick, "construct-irrelevant variance", that the assessment is too broad, and contains excess reliable

variance associated with other distinct constructs, also plays a role in CQT polygraph test results and their interpretation. This is because the outcomes of these tests can reflect other constructs, such as surprise and stress. In addition, the interpretation of CQT polygraph outcomes are flawed because they are based on a comparison of physiological responses to relevant and control questions that are non-equivalent. Thus, a larger physiological response to a control question may reflect the fact that this question focused on a particularly sensitive issue from the examinee's life history, rather than the examinee's veracity[1].

Criteria for Empirical Tests Designed to Assess Validity

CQT polygraph test results, thus, lack convergent and discriminant validity and suffer from the two major threats to construct validity. But, in terms of a court's assessment of validity, *Daubert* suggests another requirement: that the theory has been tested and an error rate estimated. The term "test", however, is ambiguous and the mere fact that data have been collected does not mean that the study is relevant and useful. Polygraph test accuracy has been tested repeatedly (cf. Iacono & Lykken, 1978; Saxe et al., 1985), but few of these empirical tests satisfy the basic requirements of a proper test. Several methodological considerations and criteria are required if criterion-related validity (or accuracy) is to be estimated properly.

Requirements for an Empirical Assessment

[1] An alternative method of psycho-physiological detection, known as the Guilty Knowledge Technique (GKT), or the Concealed Information Test (CIT) may be much less vulnerable to these threats to validity (see Lykken, 1959, 1960, 1974). The GKT is not designed to detect deception; rather, its goal is to discriminate between individuals who have knowledge about a particular event, and those who have no such knowledge. Unlike the CQT, inferences from a given physiological response pattern to knowledge about the event are based, in the GKT, on a comparison between the responses to completely equivalent questions. Furthermore, this inference is supported by Orienting Response theory (Sokolov, 1963, 1966), which postulates enhanced physiological responses to significant stimuli (see, also, Gati & Ben-Shakhar, 1990).

To estimate the validity of inferences made on the basis of a polygraph-based interrogation, one has to design a study in which those inferences are compared with a proper criterion. In the case of the polygraph – and undoubtedly in other cases as well – it is not possible, given current knowledge to design an adequate study. From a legal perspective, validity studies should meet the two criteria: (a) accuracy estimates must be generalizable to realistic circumstances and. (b) the focus is the physiological responses to the questions, rather than other information that might be available to the polygraph examiner (see Ben-Shakhar, Bar-Hillel, & Lieblich, 1986). To comply with both the scientific and the legal specifications, the following specific requirements must be met:

1. *Adequate Criterion*: A necessary requirement of any validity study is the availability of a good measure of the criterion. It is difficult to fulfill this requirement in many testing situations, but it is particularly difficult in the case of polygraph tests. They are typically conducted when incontrovertible evidence is unavailable and, thus, whether the suspect is guilty or innocent is unknown and cannot be determined with sufficient credibility. In many criminal investigations, suspects are dismissed because the evidence is insufficient, not because they are innocent. In other cases, even after charges are made and the suspects are brought to trial, the court dismisses the charges because of insufficient evidence. Even when the court makes a decision, there is no assurance that it matches the truth, and the system is tilted toward avoiding false positive decisions (i.e., finding an innocent person guilty). Several solutions for this problem are available but none is completely satisfactory.

2. *Non-Contaminated Test Results*: If the goal of a validity study is to evaluate accuracy of the psycho-physiological component of the interrogation, then inferences made on the basis of the interrogation must not be affected by any factor, other than

the subject's physiological responses. Yet, in polygraph interrogation procedures, particularly in the CQT, more information is available to the polygraph examiner than psycho-physiological data. For example, the examiner often has information about the suspect's background, impressions of police interrogators, attorney's representations, and impressions formed during the pre-test interview and during the test, itself. It is hard to differentiate between the effects of prior information and that of the specific psycho-physiological data on the inferences made. Even if charts are "blind scored", the knowledge possessed by the examiner may still have altered how the test was conducted (see, e.g., Rosenthal & Rubin, 1978). Because of this confounding, polygraph testing is a contaminated procedure and the outcome can be attributed to non-physiological, as well as the physiological information. A proper validity study must enable this confounding to be untangled.

The issue of contamination is particularly relevant to the admissibility of expert testimony, because contaminated evidence can mislead the court, even if it is valid. If the polygraph test results reflect, for example, rumors heard by the examiner, rather than the physiological responses, introducing these results in court may be a way of presenting otherwise inadmissible evidence (see Ben-Shakhar et al., 1986).

3. *Independence between Test and Criterion*: The measurement of the validation criteria must be independent of the test results, as any degree of dependence between the two might bias the validity estimates. Such dependence could exist either if the test scores directly affected the measurement of the criterion, or if the two variables were jointly affected by other factors. If a polygraph investigation yields a confession, and this confession is later used as a criterion in a validity study, independence would be violated. It is the case that confessions are often used as criteria for validating polygraphy (see Iacono & Patrick, 1991). But there are other, perhaps less

obvious cases. For example, if court decisions are used as the criterion to validate polygraphy, it must be assured that the court was not exposed to the results of the polygraph interrogation or to the conclusions made by the polygraph examiner. Furthermore, the polygraph examiner cannot have been exposed to any of the information available to the court before interrogating the subject and scoring the charts. As described below, many polygraph validity studies have not been able to establish independence between the outcome of the polygraph interrogation and the criterion.

4. *External Validity*: This term is used to describe the degree to which the results of a given study can be generalized across conditions and across subjects (see, e.g., Cook & Campbell, 1979; Saxe & Fine, 1981). Generalization is critical, because the conditions that characterize most experiments in this area are very different from the conditions of the typical criminal interrogation situation. The following factors are critical to assure external validity of a polygraph validity study:

(i) Realistic consequences. The study must be done under circumstances similar to an action of a crime and must create the types of emotional response associated with criminal investigations.

(ii) Voluntary perpetration of the crime. There is an important difference between an illegal act, or a deception, performed because the subject chose to do it, and an act performed in an experimental context because an experimenter told the subject to do it. It could be argued that subjects who deceive in an experiment are just complying with instructions of the experimenter, and consequently are not really deceiving.

(iii) Subjects should be unaware of the experimental nature of the situation: In an experimental situation, it is usually clear that the "truth" is known to the experimenter (e.g., the card chosen by the subject in a card

test design), and therefore the subject knows that ultimately the "deception" will be revealed. This is not the case in actual polygraph interrogations, and this distinction between the simulated and the real situation is interferes with external validity.

Solutions Provided by Current Research

The methodological problems that inhibit empirical efforts to estimate the accuracy of polygraph-based interrogations are difficult to overcome, but several solutions have been offered by researchers attempting to validate psycho-physiological detection methods. These solutions can be classified into two groups: (a) controlled laboratory studies which provide high levels of internal validity, but pay a heavy price in terms of external validity, and (b) actual polygraph investigations, which achieve satisfactory levels of external validity, but which are questionable with respect to all other methodological problems, especially the verification of the criterion and its independence of the outcomes of the polygraph interrogation.

The first category of studies designed to examine the validity of polygraph methods creates a situation analogues of an investigation and uses paradigms such as the “mock crime” (Saxe et al., 1983, 1985). This procedure has been employed by Raskin and his colleagues at the University of Utah (e.g., Barland & Raskin, 1975; Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Raskin & Hare, 1978), and by a small group of other researchers (e.g., Dawson, 1980). Mock crime experiments utilize true experimental designs, in that subjects are allocated randomly into "guilty" and "innocent" conditions. The typical mock crime procedure involves a simulated event, in which subjects in one group, simulating the "guilty" condition, are instructed by the experimenter to perform some act (e.g., to enter an office after the secretary has left it and take an envelope containing a $10 bill from a desk in that office). Subjects

simulating the "innocent" condition do not perform that act, and are not involved with it. In the second stage, all subjects, both "innocents" and "guilty," are interrogated by a "blind" polygraph examiner. This design answers some of the methodological concerns raised. In particular, it provides complete assurance about who is guilty and who is innocent. Also, the random assignment of subjects into conditions and the double-blind procedure guarantee independence between the criterion and the physiological data.

This design, however, is low on external validity. None of the conditions necessary to maintain external validity is satisfied by the mock crime design. Subjects who participate in these experiments are aware of the simulated nature of their task, they know that no harm will be inflicted upon them regardless of the outcome of the polygraph's interrogation, and they are not really deceiving because they are acting in accordance with the experimenter's demands. The rationale of CQT interrogation method depends on the ability of the examiner to make control questions appear more threatening for innocent suspects and relevant questions look more threatening for the guilty. Critics of CQT polygraphy (e.g., Lykken, 1998), have questioned the ability of polygraph examiners to create such a differential concern about the different types of questions for the different kinds of suspects and it is a fundamental problem with the CQT. In the mock crime situation, however, the relevant questions pose no threat to the examinees, and therefore it should be easy to formulate control questions (which, relate to real events from the subject's life history) that would be more threatening for an innocent subject than the relevant questions. Thus, such analogue studies are likely to underestimate rates of false-positive errors as compared with the more realistic situation in which real suspects are interrogated. The extent of bias in estimating the error rates in mock crime studies is unknown, but such a bias is very likely to occur

and may artificially decrease the false positive and increase the false negative rates. Therefore, despite advantages of the mock crime design, it does not allow for generalizations from the results of simulated validity studies to the real situation where suspects are interrogated regarding actual crimes.

An alternative to the experimental approach is to take polygraph charts from actual interrogations and match them with a criterion (cf. Saxe et al., 1983, 1985). Two types of criteria have been used in field studies: (a) judgments made by a panel of legal experts who are privy to all the information gathered about the case, except for the polygraph results, and (b) the use of a restricted sample of cases for whom guilt or innocence can be determined through a confession by guilty suspects (such a confession makes the confessed individual a "verified" guilty subject, while other suspects become the "verified" innocent subjects). The first approach has been used in just a few studies (e.g., Barland & Raskin, 1976; Bersh, 1969), while the confession criterion has been used frequently (e.g., Horvath, 1977; Hunter & Ash, 1973; Kleinmuntz & Szucko, 1984; Slowick & Buckley, 1975).

Both of these approaches, however, suffer from methodological problems, related to their choice of criteria. The panel criterion is problematic because (a) it is based on judgments that might be wrong and (b) the judgments made by the panel are not really independent of the judgments made by polygraphers. Although the legal experts do not have access to the actual polygraph results, dependency might be introduced because the panelists and the polygraph examiner are exposed to the same information. The use of confessions as a criterion might be even more problematic than the use of legal experts. One cannot assume that confessions are independent of the outcomes of the polygraph interrogation. Polygraph-based interrogations are designed not only to discover the truth, but also have a confession-inducing function

(see, e.g., Furedy & Liss, 1986). Polygraphers are more likely to try to induce a confession from a suspect whose chart shows clear signs of deception than from a suspect whose chart does not have such signs. Thus, a guilty suspect who showed larger responses to control questions compared to relevant questions, and therefore was classified as innocent by the polygrapher, is less likely to be included in confession-criterion studies. As a result, the sample in a typical confession-criterion validity study is biased, inasmuch as it underrepresents false negative errors (guilty subjects classified as innocent by the polygrapher). Iacono (1991) demonstrated how a polygraph examiner functioning at an overall chance level accuracy rate might accumulate a sample of polygraph records from confessed suspects with a near perfect accuracy.

A study of polygraph testing in a realistic situation that enabled complete control over the criterion, was attempted by Ginton, Daie, Elaad, and Ben-Shakhar (1982). Ginton et al.'s subjects were 21 Israeli policemen participating in a course. They were given a paper-and-pencil test that was presented as a course requirement. Beneath the answer sheet there was a hidden chemical page that received an impression of what was written on the answer sheet. The answer sheet was later separated from the rest of the pages and handed back to the subjects. The correct answer keys were then handed-out and subjects were asked to score their own tests. Subjects, thus, had an opportunity to revise their initial answers and cheat; however, the chemical copy made it possible to know exactly whether and how each subject had tampered with his answer sheet. All subjects were told that they were suspected of cheating, were offered an opportunity to take a polygraph examination, and were told that their future careers in the police force might depend on the outcome of this examination.

The deception in this situation is authentic. Both guilty and innocent subjects are truly concerned with the outcome of the interrogation and there is no question about the validity of the criterion. Unfortunately, this study highlights the difficulty inherent in validating polygraph testing. Seven of the subjects cheated, and although all 21 subjects initially agreed to take the polygraph test, one guilty subject did not show up for the examination, two subjects (one guilty and one innocent) refused to take the test, and three guilty subjects confessed just before the polygraph interrogation. The final sample, thus, included only two guilty and thirteen innocent subjects. Despite substantial efforts to meet the criteria of a proper validity study, the resulting investigation raises significant ethical problems and does not allow for proper estimates the error rates of polygraph-based decisions. It may, in fact, be impossible to conduct a proper validity study. Thus, neither are there data that can answer the *Daubert* questions about testability and known error rate, nor is it likely that such data can be developed. Unless each of the methodological requirements described above are met and a proper validity study is conducted, accuracy estimates are not useful for admissibility decisions.

Conclusions

Daubert has profound implications for our system of law and places courts in the difficult position of having to make complex scientific judgments. The present discussion of the fit, or lack thereof, between the current view of scientific validity among psychologists and behavioral scientists, and the *Daubert* criteria, make clear how complex it is to apply scientific principles in the legal framework. As demonstrated by the complexity of determining whether polygraph tests are sufficiently valid, evaluating validity is not simply a process of determining whether research is available. Nor is it a process of simply summarizing the results of research. Drawing

inferences from research requires complex judgments about the adequacy of the methodology and the degree to which plausible alternative explanations can be rejected. Moreover, scientific validation requires not only that one assess empirical evidence; rather, both the theoretical foundation and the evidence that either support or contradict it need to be jointly assessed. Thus, although *Daubert* requires courts to make scientific judgments, the two systems of thought -- legal and scientific -- are not easily aligned.

What is, perhaps, the most important disjuncture between legal and scientific thinking is that, as researchers, we cannot prove that our theories are valid. Rather, theories (and applications derived from them) are rejected by contradictory evidence. When a scientist (proponent) tells a court that a theory, and a technique derived from it, is scientifically valid, the scientist is saying that the available evidence conforms to predictions derived from the theory, and that no contradictory evidence has been produced. Another scientist (critic) who rebuts the proponent's testimony cannot prove that the technique is unreliable (in legal terminology) or invalid (in scientific terminology), rather, the critic must focus on problems with the proponent's evidence. Thus, in the case of the polygraph, proponents and critics have different burdens of proof. Courts, as referees of such disputes, are in the awkward position of having to assess these arguments and make judgments about the methodological adequacy of particular studies.

What is also difficult is that scientific criteria are relative, not absolute. Thus, the degree to which one is satisfied with the adequacy of data depends on the purposes for which the data are used. The American Psychological Association's (1985) criteria for test validity require that test validity must be established for the particular uses and populations to which one generalizes. There are cases where a given test is

valid for some uses, but not for others. For example, a study may provide valid data about a link between psycho-physiological data and deception (e.g., Furedy, Davis & Gurevich, 1988), but this is not sufficient to establish the validity of the test as evidence of criminal activity. Polygraph testing is particularly complex because it has been proposed for use in a wide variety of situations, from murder investigations to sexual assault cases, to perjury. Each of these situations may be associated with differential levels of psycho-physiological reactivity and, thus, the outcome of testing may be different and different evidence is required to establish validity in each case. As well, the consequences of an error can vary from the trivial to the profound and our code of ethics requires that we take consequences into account.

The underlying problem in making psychological research available to courts is that the fundamental issue is one of construct validity. That is, the theory from which the test (or intervention) is derived is critical. Although some have claimed that producing data demonstrating predictive (or criterion) validity is sufficient for validating the test, this is a narrow and out-dated interpretation of test theory. Unless one understands the underlying theory, knowing how the test applies in particular situations is impossible. The heart of the present critique of the polygraph is that no theory which ties deception (or any criminal activity) with physiological reactions has been formulated. Furthermore, there is "no unique physiological reaction to deception" (Saxe et al., 1983), and all physiological measures used for polygraph tests are sensitive to, and can be elicited by, a host of factors other than deception or criminal activity. If deception is not uniquely related to physiological reactions, and theory cannot explain the nature of the relationship, it is impossible to predict the conditions under which polygraph test results will be accurate or inaccurate.

Krauss and Sales (1998) note that it is unclear whether or not *Daubert*

represents a more liberal approach to expert scientific evidence than *Frye*. This ambiguity reflects the fact the lack of clarity as to whether all the requirements mentioned in *Daubert* must be met, or whether it is sufficient that the expert evidence satisfies just one of them. Our interpretation of *Daubert* suggests that although the "general acceptance" criterion was identified in *Daubert* as one of several criteria, it was not sufficient for admissibility without clear demonstrations of validity. Thus, while "general acceptance" was sufficient under *Frye*, it is no longer the case.

It may be argued that our interpretation of *Daubert* (i.e., that scientific demonstration of validity is required for admissibility of expert testimony) is too restrictive. Slobogin (1998) claims that under this interpretation a great deal of expert testimony based on the behavioral sciences should be ruled inadmissible. For example, he argued that, "Psychiatrists, psychologists and social workers who base their testimony on behavioral science information are often, at best, engaging in informed speculations, not reporting data obtained through rigorous scientific methods" (Slobogin, 1998, p. 2). Perhaps a distinction should be made between expert testimony based on scientific evidence and other technical or professional testimony based on "clinical" experience. The recent U.S. Supreme Court decision, Khumo Tire v. Carmichael (1999) indicates that *Daubert* is to be interpreted broadly and that technical and professional experts may also be subjected to scrutiny based on the *Daubert* criteria.,

What is also clear is that the relevance of the testimony is also critical. One of the problems with polygraph test evidence is that, in many situations, it speaks directly to the issue that is to be weighed by the triar of fact. Evidence that a defendant "passed" a lie detector test concerning whether or not he or she commited the act that is the focus of the trial is tantamount to deciding the case. Testimony which deals with

secondary issues, such as the mental state of the defendant may be relevant, but may speak only to the defendant's degree of responsibility . Although it seems important to apply the *Daubert* criteria to the first type of expert testimony, it may be desirable to apply less stringent criteria to the second type of evidence (e.g., general acceptance and the qualifications of the expert). Indeed, Frolik (1998) noted that expert testimony of psychiatrists, psychologists and physicians regarding mental capacity has not been challenged by the *Daubert* criteria, perhaps because it is not decisive.

There are no easy answers to the problem of the complexity of scientific logic and its non-isomorphic relationship to how courts must consider scientific evidence. Chief Justice Rehnquist's comment about the conundrum created by *Daubert* was prescient (see also, Krauss & Sales, 1998) and the directions that should be taken are not at all clear. Perhaps there is a way to establish "science courts" where specially trained judges can review scientific controversies, or to use consulting experts to the courts, as proposed by Krauss and Sales . As well, perhaps courts need to appoint experts unencumbered by being a witness for one side who can help them sort out these issues. A committee of experts may be an efficient method of assisting the courts regarding the admissibility of complex and controversial scientifically-based technologies, like the polygraph, because it may free the courts from the necessity to make admissibility decisions in each and every case, and thus save time and resources. In addition, such a committee may have better prospects of resolving the complex theoretical and methodological issues that need to be examined to assess whether the technology meets the *Daubert* criteria. Such a committee, headed by a Supreme Court Justice was nominated in Israel to examine the question of polygraph admissibility in Israeli courts (State of Israel, 1981). This advisory committee recommended that polygraph-examinations results are not to be admitted as evidence in criminal cases,

and the fact that a defendant consented or refused to be examined by a polygraph should not be brought to the court's attention (see, also, Harnon, 1982). Although the Israel Supreme Court has not taken a direct stand on the admissibility of polygraph results, polygraph tests have never been used as evidence in Israeli criminal courts (in spite of their extensive usage by the Israeli Police). The recommendations of the advisory committee are largely responsible to this state of affairs. A similar approach can be adopted for other scientific evidence of questionable validity.

It is unlikely, however, that these approaches will resolve the underlying problem. In many cases -- perhaps, the polygraph is one of them -- a minority of scientists hold very strong opinions and other scientists, no matter how skilled or objective, are unlikely to be able to resolve the controversy. Polygraph testing, in part because it assesses an area of human behavior that individuals seek to hide from others, is difficult to study and has been unusually resistent to the normal way in which scientific controversy is resolved. But because polygraph tests speak to the central issue in any legal dispute – truthfulness – the controversy is too important to be dismissed or be delayed until we have better means of answering the question.

William Marston, the progenitor of the lie detector test and a rejected witness in *Frye*, believed that a normal person could not lie without effort and without a physiological trace. Decades later, the author and physician Lewis Thomas noted that, if reports about the polygraph are true, Awhat extraordinarily good news that we are biologically designed to be truthful ... and cannot tell a lie without setting off a kind of smoke alarm somewhere deep in a dark lobule of the brain." The function of science is to examine and objectively assess such ideas. That *Daubert* requires courts to join with scientists in making such assessments is not necessarily bad. At least in the case of the admissibility of polygraph testing, despite the difficulty that judges have to

understand scientific concepts and distinguish good and bad science, courts have generally madejudgments consistent with the views of the community of scientists. The process is inefficient, and there exists the potential for error, but justice does seem to be done. Although Justice Rhenquist's fears about the demands on courts to become "amateur scientists" may have become reality, it may be a healthy development. Courts and the scientific community will need to learn how to communicate better with one another to the benefit of both.

References

AERA, APA, & NCME. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Backster, C. (1963). Polygraph professionalization through technique standardization. Law and Order, 11, 63-64.

Barland, G. H., & Raskin, D. C. (1975). An evaluation of field techniques in detection of deception. Psychophysiology, 12, 321-330.

Barland, G. H., & Raskin, D. C. (1976). Validity and reliability of polygraph examinations of criminal suspects (Report No. 76-1, Contract No. 7599-0001). Law Enforcement Assistance Administration: United States Department of Justice.

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

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

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

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

Berger, M. A. (1994). Evidentiary framework. In Federal Judicial Center (Ed.), Reference manual on scientific evidence. Rochester, NY: Lawyers Cooperative.

Berlyne, D. E. (1960). Conflict, arousal and curiosity. New York, NY: McGraw-Hill.

Bersh, P. J. (1969). A validation study of polygraph examiner judgement. Journal of Applied Psychology, 53, 399-403.

Blanck, P. D., & Berven, H. M. (1998). Evidence of disability after *Daubert.* Unpublished manuscript.

Brennan, R. L. (1992). Generalizability theory. Educational Measurement: Issues and Practice, 11, 27-34.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Committee of Concerned Scientists. (1997). United States annual report. Bayside, NY.

Commonwealth of Massachusetts v. Woodward (D. Boston, 1998), SJC-07635.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Daubert v. Merrell Dow Pharmaceuticals. 113 C. Ct. Supp. 2786 (1993).

Dawson, M. E. (1980). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. Psychophysiology, 17, 8-17.

Faigman, D. L., Kaye, D., Saks, M. J., & Sanders, J. (1997). Modern scientific evidence: The law and science of expert testimony. St. Paul, MN: West

Law.

Federal Judiciary Center. (1994). Reference manual on scientific evidence. Rochester, NY: Lawyers Cooperative.

Ford, (1996). Lies! Lies!! Lies!!!: The psychology of deceit. Washington, DC: American Psychiatric Association Press.

Frolik, L. (1998). Behavioral and social science evidence in guardianship and probate law. Psychology, Public Policy, and Law.

Frye v. United States. 293 F. Supp. 1013, 1014 (App. D.C. 1923).

Furedy, J. J., & Liss, J. (1986). Countering confessions induced by the polygraph: Of confessionals and psychological rubber hoses. The Criminal Law Quarterly, 29, 92-114.

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. Psychophysiology, 25, 683-688.

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

Ginton, A., Daie, N., Elaad, E., & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real life situation. Journal of Applied Psychology, 67, 131-137.

Harnon, E. (1982). Evidence obtained by polygraph: An Israeli perspective. The Criminal Law Review, 329-348.

Honts, C. R., & Quick, B.D. (1995). The polygraph in 1995: Progress in science and the law. North Dakota Law Review, 71 (4), 987-1020.

Horvath, F. S., & Reid, J. E. (1971). The reliability of polygraph

examiner diagnosis of truth and deception. The Journal of Criminal Law, Criminology and Police Science, 62, 276-281.

Horvath, F. S. (1977). The effect of selected variables on interpretation of polygraph records. Journal of Applied Psychology, 62, 127-136.

Hunter, F. L., & Ash, P. (1973). The accuracy of consistency of polygraph examiners' diagnoses. Journal of Police Science and Administration, 1, 370-375.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), Advances in psychophysiology (Vol. 4, pp. 1-101). London: Jessica Kingsley.

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

Iacono, W. G., & Lykken, D. T. (1998). Polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), The West companion to scientific evidence . West Publishing Co.

Katkin, E. S. (1987). Psychological assessment for decision-making: Conceptions and misconceptions. In D. R. Peterson & D. B. Fishman (Eds.) Assessment for decision (pp. 107-119). New Brunswick, NJ: Rutgers University Press.

KUMHO TIRE COMPANY v. PATRICK CARMICHAEL, etc.(1999).No. 97-1709.

Kimmel, H. D., Van Olst, E. H., & Orlebeke, J. E. (1979). The orienting reflex in humans. New York, NY: Erlbaum.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. Journal of Applied Psychology,

73, 291-302.

Kleinmuntz, B., & Szucko, J. J. (1984). Lie detection in ancient and modern times: A call for contemporary scientific study. American Psychologist, 39, 766-776.

Krauss, D. A., & Sales, B. D. (1998). The problem of "helpfulness" in applying Daubert to expert testimony: Child custody determinations in family law as an exemplar. Psychology, Public Policy, and Law.

Loftus, E. F. (1979). Eyewitness testimony. Cambridge, MA: Harvard University Press.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lykken, D. T. (1978). The psychopath and the lie detector. Psychophysiology, 15, 137-142.

Lykken, D. T. (1979). The detection of deception. Psychological Bulletin, 86, 47-53.

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

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. Journal of Experimental Psychology, 2, 117-163.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Patrick, C. J., & Iacono, W. G. (1991). Validity of the control question

polygraph test: The problem of sampling bias. Journal of Applied Psychology, 76, 229-238.

Podlesny, J. A., & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. Psychophysiology, 15, 344-359.

Penrod, S. D., Fulero, S. M., & Cutler, B. L. (1995). Expert psychological testimony on eyewitness reliability before and after Daubert: The state of the law and the science. Behavioral Sciences and the Law, 13, 229-259.

Raskin, D. C., & Hare, R. D. (1978). Psychopathology and detection of deception in a prison population. Psychophysiology, 15, 126-136.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), Modern scientific evidence: The law and science of expert testimony. St. Paul, MN: West Law.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1998). Polygraph techniques. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), The West companion to scientific evidence. West Publishing Co.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral Brain Sciences, 1, 377-415.

Saxe, L. (1991a). Lying: Thoughts of an applied social psychologist. American Psychologist, 46, 409-415.

Saxe, L. (1991b). Science and the CQT polygraph: A theoretical critique. Integrative Physiological and Behavioral Science, 26, 223-231.

Saxe, L. (1994). Detection of deception. Current Directions in Psychological Science, 3, 69-73.

Saxe, L., Dougherty, D., & Cross, T. P. (1983). Scientific validity of polygraph testing (OTA-TM-H-15). (Report for the U.S. Congress Office of Technology Assessment). Washington, DC: U. S. Government Printing Office.

Saxe, L., Dougherty, D., & Cross, T. P. (1985). The validity of polygraph testing: Scientific analysis and public controversy. American Psychologist, 40, 355-366.

Saxe, L., & Fine, M. (1981). Social experiments methods for design and evaluation. Beverly Hills: Sage Publications.

Selye, H. (1976). The stress of life. New York, NY: McGraw-Hill.

Slobigin, C. (1998). The admissibility of behavioral science information in criminal trials: From primitivism to Daubert to voice. Psychology, Public Policy, and Law.

Slowick, S. M., & Buckley, J. P. (1975). Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure and GSR recordings. Journal of Police Science and Administration, 3, 305-309.

Sokolov, E. N. (1963). Perception and the conditioned reflex. New York, NY: Macmillan.

Sokolov, E. N. (1966). Orienting reflex as information regulator. In A. Leontyev, A. Luria, & A. Smirnov (Eds.), Psychological research in the U.S.S.R. (pp. 334-360). Moscow: Progress Publishers.

State of Israel. (1981). A report regarding the uses of the polygraph. Jerusalem: Ministry of Justice (in Hebrew).

United States v. Frank Javier Cordoba, 158 (D. California, 1998), aff'd, SA CR 95-39-GLT[SF].

U.S. v Pitner et al., (D. Washington, 1997), Dkt. No, 211 in CR 96-500C.

United States v. Scheffer. 118 S. Ct. Supp. 1261 (D. Washington, 1998), aff'd, USCA Dkt. No. 95-0521/AF (US Court of Appeals for the Armed Forces).

Author Note

Leonard Saxe, Heller School; Gershon Ben-Shakhar, Department of Psychology.

We thank Maya Bar-Hillel, Michal Beller, Theodore Cross, Eliahu Harnon, Mordechai Kremnitzer, Sonny Kugelmass, and David Weisburd for their helpful comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Prof. Leonard Saxe, Heller School, MS 35, Brandeis University, Waltham, MA 02454-9110. Electronic mail may be sent via Internet to saxe@brandeis.edu.