Polygraph

VOLUME 33 2004 NUMBER 3

Contents

Published Quarterly

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Directed Lie Comparison Questions in Polygraph Examinations: History and Methodology

Paul M. Menges

Abstract

Directed Lie Comparison (DLC) questions were used as early as the late 1960s by a small number of government polygraph examiners in a variety of polygraph testing formats. The Test for Espionage and Sabotage (TES), used extensively by some government agencies since 1993, utilizes DLCs. However, prior to the advent of TES, the number of examinations conducted using DLC questions was relatively small. Since the explosion in the use of DLCs, primarily as a result of the TES technique, concern has been voiced in some quarters concerning the accuracy, validity, and the susceptibility to countermeasures of DLC questions. This paper is not a critique or endorsement of the TES format but is intended to address the broader issue involving the use of DLCs in any format. It provides additional insight into the development of the DLC technique and suggests the DLC technique as used by its original designers may have been more sophisticated than many examiners realize. In fact, those who developed the technique described it as unique and worthy of additional emphasis and caution when training examiners in its use (Paul C. Dubiel, personal communication, February 3, 2003; Fuse, 1982; and Lloyd H. Hitchcock, personal communication, February 5, 2003). This paper does not suggest the original methodology was ideal or that DLCs must be used exactly as originally designed. However, it does suggest the need to reconsider some aspects of current methodology for use and evaluation of DLCs. While the use of DLCs may not hold the answer for all test requirements, it is a valuable tool with a variety of applications.

Background

Empirical data compiled since 1987 demonstrate the accuracy of polygraph examinations using DLCs (Alloway & Honts, 2002; Barland, 1981; Department of Defense Polygraph Institute [DoDPI], 1997; DoDPI, 1998; Honts & Raskin, 1988; Horowitz, Kircher, Honts & Raskin, 1997; Kircher, Packard, Bell & Bernhardt, 2001; Reed, 1994). No single polygraph technique provides the answer in every situation; however, DLCs provide several advantages over PLCs in some situations. They are less intrusive, more easily standardized, and avoid some of the difficulties experienced in developing PLCs during repeated testing (Barland, 1981; Fuse, 1982; DoDPI, 1997; Horowitz, et al., 1997; Kircher et al., 2001). DLCs also provide a basis for a more standardized, objective manual evaluation than is possible with a Relevant-Irrelevant (RI) question polygraph examination. These attributes are meaningful and significant to any examiner who has had the experience of extensive and repeated testing of any examinee. When properly used, DLCs are particularly beneficial when conducting routine screening of individuals not suspected of any wrongdoing.

Some of the earliest discussions of DLCs are found in unpublished papers written by former examiners in the U.S. Army Military Intelligence (MI) Polygraph Program. Louis Fuse (1982), a Chief Warrant Officer (Retired) and former polygraph examiner with the U.S. Army MI Polygraph Program explained the development, rationale, and use of DLCs by MI examiners during a presentation to the Federal Interagency Polygraph Seminar (FIPS) at Quantico, Virginia, in June 1982. Fuse's 1982 unpublished manuscript (*Directed Lie Control Testing Technique*) is often cited as one of, if not the original introduction to the DLC technique in the polygraph literature. William A. Stallsmith, another former MI examiner updated and revised Fuse's manuscript in 1987, using it as a training document for MI examiners.

Following Fuse's introduction of the DLC concept, several researchers and examiners experimented with and reported their findings regarding DLC question use, validity, and accuracy. The resulting

discussions and occasional debate regarding the best method of introducing DLCs, whether DLCs required inter-chart stimulation, or whether DLCs should be evaluated in the same manner as PLCs, focused deserved attention on a little understood and little researched technique (Abrams, 1991; Abrams, 1999; Honts & Gordon, 1998; Honts, 1999; Honts, Raskin, Amato, Gordon & Devitt, 2000; Matte, 1998; Matte, 2000 and Matte & Reuss, 1999).

Some of the original DLC rules or concepts conflict with what has been researched or developed by field examiners since the general introduction of the technique to the larger polygraph community. For example, while MI examiners rarely employed DLCs in specific issue type examinations, most of the research cited above reported high accuracy while using DLCs in mock crime situations. Additionally, Fuse (1982) suggested the occasional need for mild inter-chart stimulation of an examinee who did not appear to be responding to the DLCs. However, he warned of the possible repercussions of overemphasizing the DLCs. Indeed, by the late 1980s, MI examiners were encouraged to provide minimal stimulation, if any, between charts and if done, it was to be balanced. A general statement to the effect of "just answer all questions truthfully, except those to which you are lying" or "it's clear to see where you're lying" was considered appropriate. However, these procedures, as well as those used for evaluation and described later in this paper were based on experience and anecdotal evidence.

It is also important to recognize that the DLC technique designed by the examiners mentioned in this paper differed considerably from all other techniques used at the time. The DLC technique employed by MI examiners was believed to require more finesse and pretest preparation for *balance* than that normally required when using PLCs or other formats such as a "Yes Test" (Reid & Inbau, 1977), "Yes-No" (Golden, 1969), or "Positive Control Technique" (Reali, 1978). In the "Yes Test" the subject was told to answer each question affirmatively. In the other formats the examinee was told the questions would be repeated and he or she was expected to answer truthfully one time and lie the next time the question was asked; or vice versa, depending on which format was used.

The original DLC technique encompassed a detailed process wherein the examinee received an explanation and demonstration of the physiological activity being monitored during testing. Subjects were told that when one lied there was generally a physiological response and when subjects were truthful, they generally did not exhibit consistent, significant physiological responses to questions during testing. The demonstration, an acquaintance test described later in this paper, usually made the method of operation clear to an examinee. The collection of biographic information and minor details of an examinee's status, employment, medical situation (to determine suitability for testing), was deemed important to balance the discussion, develop rapport, and develop minor details suitable for tailoring DLCs to the individual examinee later in the process. The specific DLCs used involved personal involvement and encompassed minor transgressions that most examinees could acknowledge. In fact, in this process the comparison questions became known lies. The physiological responses to the DLCs were presumed to be the result of cognitive awareness and similar to responses evidenced in known number acquaintance tests, where presumably, little to no fear of detection existed.

A case for DLCs...a little history

Chief Warrant Officer Paul C. Dubiel, a U.S. Army examiner, now retired, authored several unpublished and undated papers used in instructing MI examiners in the 1970s and 1980s (F. D. Clifton, personal communication, July 13, 2001 and P.C. Dubiel, personal communication, February 3, 2003). Dubiel was Chief Examiner with the 902nd Military Intelligence Group (MIG) in Arlington, VA, prior to his retirement in 1973. He was introduced to the concept of DLCs, then described as "weak controls," by Lloyd H. ("Rusty") Hitchcock, another former MI examiner, retired Warrant Officer, and World War II veteran.

In the late 1960's, Dubiel was assigned as Chief Examiner with the 66th MIG in

Europe and met Hitchcock, who was assigned as a civilian Assistant Operations Officer. Dubiel described Hitchcock as his "intelligence polygraph mentor," adding that he (Hitchcock) explained and demonstrated a polygraph technique he developed to accommodate the nuances of intelligence polygraph testing (P. C. Dubiel, personal communication, February 3, 2003). At that time, according to Dubiel, "Our only polygraph training was in the criminal field and it was far insufficient to successfully meet the intelligence requirements levied upon us and by necessity we had to improvise to meet the responsibilities imposed upon us as the lives of our fellow military members were often at stake" (P. C. Dubiel, personal communication, February 3, 2003).

Hitchcock's work in the area of DLCs was driven by his distaste for some of the polygraph techniques then in use, such as stimulation tests which used what he termed as "card tricks" or some sleight-of-hand to convince an examinee that the examiner was able to correctly discern an examinee's selection of a number or card using the polygraph. He felt strongly that the polygraph, in the hands of competent examiners, provided a valuable tool in the process of detecting deception and verifying information. Hitchcock was adamant that examiners should be as truthful as possible with examinees to obtain quality results and that "honesty begets honesty" (L. H. Hitchcock, personal communication, February 5, 2003 and P. C. Dubiel, personal communication, February 3, 2003). He strongly believed in the benefits of a polygraph technique that stressed an upfront approach with the examinee. Hitchcock was concerned about the nuances involved in the repeated testing of sources, many of whom could best be described as unsavory characters. In these cases, caution was necessary to ensure traditional comparison questions were not relevant. DLCs alleviated this concern. These thoughts and observations provided the foundation for his belief in and utilization of DLCs.

While serving in the Republic of Vietnam during 1970 and 1971, Dubiel trained Fuse and Mary V. Bender, also a former Army Warrant Officer and examiner, and continued to develop the DLC technique (L. Fuse, personal communication, May 15, 2001, F. D. Clifton, personal communication, January 10, 2003 and P. C. Dubiel, personal communication February 3, 2003). As the procedure evolved and was taught to other MI examiners, the DLC technique was strictly structured and to a degree, scripted, to ensure the proper preparation of an examinee, including the presentation and use of the Acquaintance Test (ACQT) as explained by Fuse (1982).

A basic premise behind the development of the DLCs was the need to develop and vet information in a variety of testing situations involving source reporting and applicant testing where no issue was known to exist. Such examinations often required repeated and extended testing, as information was developed, then verified through further testing. DLCs met this challenge by allowing for a less confrontational environment and were thought to be ideal for non-suspect type screening examinations. A conservatively applied objective manual analysis implemented in the 1980's was thought to have added the strength of objectivity to the technique.

Original Design

The DLC Technique used by MI examiners involved a detailed pretest discussion wherein the examiner developed rapport with the examinee. The examiner also described physiology as it pertained to polygraph and its relationship to responses that occurred when subjects lied, vice the absence of consistent and significant physiological responses when subjects were truthful. The pretest interview, which included an explanation of what to expect during the examination and a thorough discussion of relevant issues, was then followed by an ACQT, described by Barland (1981) and Fuse (1982). The ACQT was an unknown number test designed by Hitchcock to support the explanation of procedures provided in the pretest interview and to clearly demonstrate to the subject that the process was accurate and above board (L. H. Hitchcock, personal communication, February 5, 2003 and P. C. Dubiel, personal communication, February 3, 2003). It contained unique aspects and questions in addition to the standard numerical sequence used to question an

examinee regarding a number that had been selected. A "Did you lie…" question and "Now answer truthfully…" instruction, described by Barland (1981) and Fuse (1982) were utilized to increase the accuracy and professional conduct of the ACQT. The value and importance of the ACQT were believed to be so critical to the preparation of an examinee for a DLC examination, that it was required when using DLCs. Additionally, when an ACQT was conducted wherein the key could not be clearly and correctly identified, the examiner would not make a selection. In such cases, absent some logical explanation for the results, e.g., the examinee attaching some previously unknown significance to another number, followed by a successful ACQT, the examinee would not normally be tested on that date. The belief was that the examinee, for whatever reason, e.g., fatigue, medication or other outside issues, was an unsuitable candidate for testing at that time. Note that the term ACQT was used by MI examiners to denote the unknown number test required to be used with their DLC examination and separated it from the Stimulation Test, a known number test, taught by many polygraph schools at the time. Today, many polygraph schools refer to the known number test as an ACQT, as well.

With MI examiners reporting successes developing information and resolving examinations, DLCs soon became the program's preferred method to use when testing multiple issues in routine initial or periodic screening intelligence examinations. Examiners who became experienced with DLCs in this type of screening believed that PLCs frequently overpowered relevant issues when testing in intelligence field situations where no specific issue or allegation was present. Additionally, as Barland (1981) pointed out, the use of DLCs alleviated concerns about the difficulty of developing non-relevant PLCs when testing intelligence sources. MI examiners believed that the use of DLCs brought a degree of objectivity to routine screening that was not present with RI type testing, while not creating too powerful an emotional distracter, as was believed might be the case with PLCs in a general screening type examination. Like the RI type examination, the pretest discussion and preparation were extremely important in order to provide the proper balance and setting for the examination. If examinees responded physiologically, consistently and significantly to relevant issues during testing, further discussion of the relevant topics, followed by confirmatory testing was the solution of choice.

DLCs were constructed to ensure personal involvement and they were normally related to minor transgressions that almost any examinee could acknowledge (Fuse, 1982). In fact, in a study of comparison questions, Horowitz, et al., (1997) found that DLCs where personal involvement was present, e.g., Did you ever…?, produced greater accuracy than DLCs with no personal involvement, e.g., Are the lights on? DLCs were specifically not intended to be embarrassing or significantly emotion-evoking.

Hitchcock, Dubiel and Fuse (1982) stressed the critical nature of the pretest interview when using DLCs (P. C. Dubiel, personal communication, February 3, 2003 and L. H. Hitchcock, February 5, 2003). As noted previously, DLCs were not intended to provide a strong, emotion-evoking distraction for the truthful examinee. Therefore, examiners using DLCs were taught to balance the relevant issues with a pretest that exuded competence, professionalism, and confidence, thereby easing the concerns of the truthful examinee. A balanced, non-confrontational environment was created during the pretest interview wherein the examinee was prepared for the examination through what Fuse called "the physiology pretest" (Fuse, 1982). This allowed for the maintenance of an environment wherein information could continue to be developed and further testing conducted to resolve issues, during repeated and extended testing of intelligence sources. The pretest discussion described by Fuse (1982) was used by MI examiners until approximately 1993 when many in the government community began using DLCs in the manner developed by Reed (1994) for TES examinations.

MI examiners using the required DLC pretest described by Fuse (1982) frequently heard examiners unfamiliar with DLCs comment derisively about the lengthy pretest requirement. Some analysis of Fuse's

instructions for formulating DLCs and the related pretest postulated it was too complex and time-consuming (Honts & Gordon, 1998). The significance of this difference, if any, is unknown. It is believed that most DLC examinations conducted in recent years, both in the research cited earlier and in field situations, have not used the lengthier and more detailed pretest discussions of physiology practiced by the early DLC examiners. However, it is clear the intent of the pretest described by Fuse (1982) was to ensure balance through a competent, professional pretest presentation.

The ACQT was an integral part of the DLC package designed by Hitchcock and closely tied to the DLC introduction. Examiners were taught to explain the results of the ACQT, making clear the logical connection between cooperation, telling the truth, and what occurred when the examinee lied on the ACQT. It was believed that a proper foundation for the DLCs, made it unnecessary, ill advised, and possibly unethical, to add extra weight to the DLCs by providing special instructions to ensure response at the DLCs. When using DLCs, a subject should be told to remember and recognize the particular deed in question or item they are lying about. However, the manner in which this occurs was thought to be important. It was believed that any imbalance in presentation of relevant questions and DLCs could render the results inconclusive at best or a false positive or false negative at worst. This was a precarious line to tread for the novice examiner. As a result, only experienced examiners or examiners being trained and monitored by more senior examiners in the MI Polygraph Program were allowed to use the technique.

The importance of the ACQT to DLC testing was confirmed by Kircher et al. (2001), who found the use of a stimulation test with positive feedback to the subject positively impacted the accuracy of the DLC examination. The stimulation test used by Kircher et al. (2001) was a known number test. One would expect a successful, professionally conducted unknown number ACQT to have at least as much of an impact on an examination. The majority of the DLC research previously cited utilized known number ACQTs. The explanation of and emphasis placed on the unknown number ACQT by MI examiners is provided to create an understanding of their DLC methodology.

It appears from discussions with some examiners that they believe when an initial series using DLCs is inconclusive the answer is to switch to another technique such as one using PLCs. The implication of course, is that the DLC technique did not work and another technique is required to resolve the examination. Absent some posttest discussion that might result in an explanation for the physiological responses or a modification to the relevant question, this methodology appears to invite the possibility of increased errors, possibly allowing untruthful subjects to pass the examination. Most examiners know that to ignore consistent and significant physiological responses to relevant questions is a recipe for disaster. Early DLC examiners frequently warned of this dilemma (L. H. Hitchcock, personal communication, February 5, 2003, F. D. Clifton, personal communication, July 13, 2001 and P. C. Dubiel, personal communication, February 3, 2003).

A discussion of DLC practices...

Fuse (1982) described DLC question introduction wherein the examinee was asked to recall an incident or event to be used in the DLC question. This basic instruction is occasionally taken a step further by some examiners who feel that comparison questions must be strong, emotion evoking questions. If an examiner comes from a background of using PLCs, and one wherein those PLCs were expected to balance an examination involving a specific allegation, this belief is understandable. In routine screening type examinations, where no allegation exists, no previous interrogation has taken place, and no suspect is being tested, the atmosphere should be less emotional and less threatening to an examinee. If DLCs are “pushed” or made too strong, one might expect the resulting tracings to reflect exaggerated physiological activity similar to what might occur when countermeasures are being employed. In fact, MI examiners being indoctrinated in the DLC technique were warned early on in their internship about the effects of what was affectionately termed “DLC Breathing.” Such

distortion was sufficient basis for an inconclusive evaluation, requiring the retesting of a subject. MI examiners learning about DLCs were taught to have the examinee recall a specific incident related to the DLC issue in order to have something specific about which they were lying. When questions were presented during the test, the examinee was expected to recognize the question and answer in a timely fashion. Often, warning an examinee about the effects of mis-answering during testing was enough to cause the truthful examinee to pay attention and cooperate fully. In fact, a side effect and possible benefit of the DLC may be that the examinee must pay attention and answer in a fashion not normally expected of a polygraph examinee, e.g., being directed to lie to any question on a test. This requirement places a burden on the subject not always recognized by examiners and one that can be of great value in detecting and discerning some deliberate actions by subjects attempting to manipulate the test.

As previously noted, a common misconception surrounding DLCs involves the belief that the DLC question must evoke strong emotions. Some examiners and researchers comment about the emotionality of a DLC question to compare against the emotionality of a relevant question dealing with espionage in a routine screening examination. Unfortunately they are not alone in their misunderstanding of the DLC technique, at least as it was originally designed. DLCs were not originally intended to provide a strict comparison between a specific relevant issue (allegation) and an emotion evoking comparison issue. DLCs were intended to provide the examinee an opportunity to demonstrate some physiological capability, while providing additional impetus for the examinee to pay attention throughout the examination (Fuse, 1982 and F. D. Clifton, personal communications, July 13, 2001 and 10 January 2003). Evaluation of the physiological data included a conservative analysis of overall physiological activity at the relevant issues. Indeed, an undated training document (ca. 1975) used by MI examiners in the early days of DLC usage, points out that "Responses to weak control questions are not necessarily compared with responses to relevant questions for greater magnitude, per se, but are used to ensure that the examinee is still responding normally in deception" (F. D. Clifton, personal communication, July 13, 2001).

Some examiners believe for a DLC to be effective, the examinee must concentrate on and visualize the event in question for an extended period prior to lying to the DLC during the exam. This is a misinterpretation of the original DLC instructions. During question review an examinee must be asked to recall or bring to mind a specific event related to the DLC issue so he or she will have something specific to which he or she is lying. When the question is asked and answered during the test, the examinee must recognize the question and know that he or she is lying. There is a fine line between this type of instruction and asking the examinee to continue to visualize the event for a time before answering.

A major concern with DLCs involves the possibility of an increase in false negative findings if too much emphasis is placed on the DLCs by the examiner. Indeed, Fuse warned that only skilled, experienced examiners should use the technique (Louis Fuse, personal communication, May 15, 2001). He recognized the ease with which the technique could be unintentionally misused by ethical and well-intentioned examiners. Abrams (1991) also warned, "there are many ways of tilting the delicate balance between control and relevant questions" (p. 31).

When some examiners evaluate a DLC type exam as No Opinion or Inconclusive, they immediately believe the problem is that the DLCs are weak or the examinee is not concentrating sufficiently on the DLCs. In fact, in most cases, the problem is not with the DLC but a result of consistent responses to one or more of the relevant issues on the examination. To simply advise the examinee to concentrate on and visualize specific acts when answering the DLCs, and collect another chart, may increase the risk of a false negative result. To illustrate this point, an example of what would have been considered improper use and introduction of a DLC in the MI program, can be related here. On one occasion, an examiner was observed and heard introducing a DLC in substantially the following manner:

> "Now for the next question, I want you to answer me truthfully but don't provide any details. Did you ever say anything derogatory about someone behind their back?" After the examinee answered positively, the examiner continued, and pointed out that this would be a question used on the test but that the examinee would be answering "No," and thereby lying on the exam. "In addition, there will be three keys you will need to remember. (1) Do you recall what it was that you said? (2) Do you recall when you said that? And (3) Do you recall about whom you were talking?" When the examinee answered affirmatively to each of the three keys, he was told to listen to the question as it was posed during the exam. He was then supposed to "play back each key" in his mind, before answering in the negative. At that point, the examinee pointed out that such a response would require a delay of several seconds. The examiner confirmed that by saying that it would take 5-7 seconds for this processing before the answer.

This example resulted in answers to DLCs during testing that were delayed anywhere from 10-13 seconds. During that period, the respiration patterns were exaggerated to the point of distortion, the electrodermal channel displayed exaggerated complexity, and most often the cardio channel reflected similar exaggerated responses. If one accepted the physiological tracings as legitimate and utilized a strict numerical spot analysis, an NDI or NSR opinion was all but guaranteed. This type of stimulation at any comparison question is unnecessary.

Another concern voiced by many examiners involved with DLCs since the mid 1990s relates to the susceptibility of DLCs to polygraph countermeasures. Along with increased use of polygraph in screening applications has come an increased awareness of countermeasures employed in polygraph examinations. The concept of telling an examinee to lie to a specific question, while thinking about a specific act, has drawn criticism from some within the polygraph community. It is well known that mental activity can be used as a countermeasure (Honts, Raskin, & Kircher, 1994). While anecdotal evidence suggests that examiners today are much more adept at identifying countermeasures than previous research indicated, the fact that an examiner instructs an examinee to lie about an incident they can recall, and the manner in which the examinee is instructed can be cause for concern (P.C. Dubiel, personal communication, February 4, 2003; Fuse, 1982 and Louis Fuse, personal communication, May 15, 2001). It is very easy to cause an examinee to generate large physiological responses by the manner in which the examiner introduces the DLC during the examination. There is a fine line separating proper DLC instructions to an examinee and improper instructions, which could cause exaggerated physiological tracings.

Interestingly, early DLC instructions and training papers presented to MI examiners noted that a skilled, experienced examiner using DLCs had an advantage in discerning some deliberate attempts by subjects trying to manipulate the test. Dubiel believed that many guilty examinees would push their reactions to the "weak control questions," as he called the DLCs, hoping to have the examiner believe that such reactions were legitimate physiological responses. He felt the experienced examiner could easily discern legitimate reactions from what he called "pushed" reactions in the polygraph tracings. He also advised caution when reactions to DLCs were similar to the reactions to relevant questions. In such cases, the subject was considered to be having some problem with the relevant question and an interview of the subject was in order (P. C. Dubiel, personal communication, February 4, 2003).

Indeed, a basic polygraph tenet is that it is very difficult for untruthful examinees to inhibit physiological responses to relevant issues. Lykken (1998), while addressing polygraph accuracy, validity, and methods to defeat the polygraph, acknowledged that it was probably impossible for most people to inhibit responses to relevant questions.

Analysis

Objective manual scoring of DLC type examinations should be accomplished with a conservative numerical evaluation. As already noted, DLCs occasionally result in tracings which appear to exhibit exaggerated responses. A conservative analysis allows the examiner to recognize consistent, significant, and timely physiological responses at relevant questions and not be blinded by large positive scores at spots that may not be deserving of such numerical analysis.

In fact, the conservative nature of the DLC evaluation process used by MI was not unlike that described and suggested by Krapohl, Stern and Bronkema (2003). In one of their examples, a careful analysis of pneumograph patterns, which were very similar at all comparison questions, rendered large positive scores. The other two channels tended toward negative scores but resulted in the sum of all three channels yielding positive totals. In this case, they questioned whether judgment should be withheld pending closer scrutiny and perhaps further testing (Krapohl et al., 2003).

Initially, DLC examinations were evaluated using a global analysis to determine whether there was consistent, significant, and specific response to relevant questions included in the exam. Fuse noted that objective numerical analysis of DLCs was not instituted until the early 1980s. At that time numerical scoring was devised primarily as an aid for new examiners (Louis Fuse, personal communication, May 15, 2001). When a DLC question test was numerically evaluated, a 5-position scale using a range of +2 to –2 was allowed. Examiners were trained to be alert for consistent, significant, and timely response to relevant questions.

Examiners experienced with DLCs have known for some time that critical attention must be paid to the respiration pattern at DLCs in order to determine whether they are diagnostic, and if so, to what degree. Even after a numerical analysis was implemented with the DLC technique in the MI program, it was a conservative analysis to the extent that if the respiration at the DLCs was distorted, that channel was not used. At that point a determination had to be made whether the other channels provided diagnostic value or were so affected as to provide no value.

In the 1980's, MI examiners conducting technical reviews of field examinations wherein DLCs were used commented critically to examiners who submitted charts for review that displayed what came to be called "DLC Breathing." "DLC Breathing" can be defined as pneumograph patterns, frequently with exaggerated physiological tracings, usually displaying consistent distortions at DLCs across a chart, to the extent that one could discern the DLCs from the pneumograph pattern alone. In fact, such tracings would best be described as "too good to be true," a phrase used repeatedly since approximately 1995 by Gordon H. Barland, Ph.D., when identifying physiological tracings created by polygraph countermeasures attempts (Barland, 2003). When this is observed or suspected, it would behoove the examiner to question the diagnostic value of the pneumograph tracing and consider the remaining channels of physiological data. Interestingly, the critical treatment of "DLC Breathing" by those using DLCs as originally designed appears in hindsight to have been on target.

A study by Kircher et al. (2001), suggested that new procedures to evaluate DLC respiration responses are required. These findings support a previous finding by Horowitz et al., (1997) and a basic point of this paper. DLC respiration patterns should not be evaluated in the same manner as those exhibited in PLC testing.

Kircher et al.'s (2001) finding that truthful subjects tend to attempt to enhance their respiratory responses at the DLCs confirms the experiences of long-time DLC examiners. It also validates the treatment of "DLC Breathing" by examiners who refused to accept the contrived tracings as normal physiology. Truthful subjects exhibiting such tracings normally stopped the enhanced respiratory responses when counseled in a professional manner consistent with a non-threatening, non-accusatory, non-confrontational presentation. If an examinee continued the exaggerated respiration and responded relevant issues in a consistent,

significant, and timely manner, the examiner was obligated to address the relevant issues in further discussion, followed by additional testing, if appropriate.

The major difference in test data analysis between some examiners currently using DLCs and examiners using the earlier DLC methods is the manner in which charts were globally analyzed prior to and after numerical scoring was applied. In the original design, if respiration patterns at the DLCs looked "too good to be true" or extremely exaggerated, little weight was given unless the relevant question displayed absolutely no response. In these cases, appropriate instructions were provided to the examinee and the test re-done. Additionally, if a relevant question displayed any significant and consistent response, that question was closely scrutinized. If answers were delayed to the DLCs and clear, continuing response was evident in the EDA channel at the DLC, frequently almost to the next question, caution was exercised in evaluating that parameter also. Similar caution was exercised in the cardio channel when the respiration pattern was exaggerated and clearly affected that channel. In all cases where questionable DLC activity was evidenced, the focus was shifted to the relevant question to determine significance and consistency of response.

As previously noted, examiners taught to use DLCs in the original design employed the technique over two, and later three charts, with single presentations of the relevant questions, most often bracketed by DLCs. A scoring rule used by examiners in the late 1980s and one designed as an indicator of relevant response was the ability of a relevant spot to garner two positive spot totals during three presentations over three charts. Regardless of spot score, if a relevant question did not result in positive totals at two scoring spots, normally two charts out of three, the question would be considered inconclusive at best. If a relevant question stood out as consistent and significant, the DLC responses were closely scrutinized prior to application of a positive numerical analysis. This rule played a role in making the DLC decision process a conservative one and one well suited to initial screening in non-suspect situations.

Conclusion

Research, observation, and anecdotal experiences demonstrate that respiration patterns at DLCs frequently display atypical physiological tracings (Horowitz et al., 1997; Kircher et al., 2001). This is possibly caused by the mental activity occurring when the subject is attempting to do exactly what the examiner requested during the pretest instruction and question review, e.g., mental visualization, delayed answering, etc. It is this very activity that requires a delicate balance by the examiner when explaining DLC instructions to an examinee. If DLCs are emphasized as described earlier, the distortion evident in DLC respiration patterns, if not recognized as atypical, may be so great that it causes a positive evaluation in cases not rightly deserved, hence higher false negative results.

In most polygraph screening applications such a result would be disastrous. One solution to non-suspect, routine applicant or counterintelligence scope screening polygraph examinations is to initiate routine screening polygraph examinations with a technique that is extremely sensitive to relevant response, not intrusive to examinees, and more scientifically defensible. To use a medical model, such as that suggested by Krapohl & Stern (2003), an initial screening tool must be extremely conservative to preclude a false negative result that adversely impacts national security or an agency's ability to accomplish its mission. Those examinations not resolved in the initial phase can be expected to be resolved through more specific follow-up testing.

Dr. Charles Honts (personal communication, July 5, 2004) suggested that such a conservative approach will result in an increase in false positive results. This may be the case if based on the assumption that both negative (NDI) and positive (DI) results are rendered in the screening phase of the examination. Donald J. Krapohl (personal communication, August 20, 2004) suggested that, "If a medical model can be applied to the phased testing, perhaps the concerns about false positives can be mitigated or reduced. While negative results are permitted in the preliminary stage if warranted, reactions to the

relevant questions trigger a subsequent and more focused testing process, not an immediate DI decision. In these latter instances, tests with better specificity are brought to bear to help distinguish between those individuals who are withholding reportable information from those whose responses were simply random or caused by concerns about non-reportable matters." DLCs are integral to this process. They are part of a sensitive screening tool that can be used over a series of examinations to address problem areas, while allowing for an entirely professional, above-board environment throughout the screening process.

In the final analysis, if an examiner recognizes consistent, significant, and specific responses to relevant questions, and is vigilant for extravagant (too good to be true) physiological activity at the DLCs, this technique can be effectively employed. There should be little doubt about the legitimacy of a response before positive numerical analysis is applied. However, currently accepted DLC methodology requiring spot evaluation using the same rules that apply to PLCs must be reevaluated in order to meet this challenge. Research scrutinizing numerical cut-off thresholds and new treatment of respiration tracings when using DLCs may provide even more accurate results than already evidenced by research.

Personal experiences and observations lead to the belief that most of the difficulties and concerns with DLCs lie in the experience level of the examiner, the level of confidence the examiner possesses in the DLC technique, and his or her ability to recognize consistent, significant, and specific physiological response. As noted earlier, examiners inexperienced with and lacking confidence in DLCs, when faced with an unresolved test, may be inclined to quickly change techniques vice acknowledging the responsiveness of the relevant issue. DLCs provide a tremendous tool to the experienced examiner but they also present a challenge. The technique can be easily misunderstood, misapplied, and misused. Training, experience, mentoring, and quality supervision are critical to ensuring that examiners are prepared to utilize DLCs in appropriate test formats.

Acknowledgements

The author is a federal examiner, trained in the use of DLCs in the MI Polygraph Program, in 1986. Paul C. Dubiel, Louis S. Fuse, F. Don Clifton, and Lloyd H. Hitchcock, all former MI examiners, provided invaluable assistance with their personal insights and experiences with DLCs. Gordon H. Barland, Ph.D., Andrew B. Dollins, Ph.D., Charles R. Honts, Ph.D., Frank S. Horvath, Ph.D., Donald J. Krapohl, Stuart M. Senter, Ph.D., and Dan V. Weatherman, also greatly assisted the author with their diligent review and critique of this paper.

References

Abrams, S. (1991). The directed lie control question. *Polygraph,* 20, 26-31.

Abrams, S. (1999). A response to Honts on the issue of the discussion of questions between charts. *Polygraph,* 28 (3), 223-229.

Alloway, W.R., & Honts, C.R. (2002, April). *An information countermeasures has no effect on the validity of the Test for Espionage and Sabotage (TES).* Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Park City, UT.

Barland, G.H. (1981). *A validation and reliability study of counterintelligence screening tests.* Unpublished manuscript, Security Support Battalion, 902d Military Intelligence Group, Fort George G. Meade, MD.

Barland, G.H. (2003, August). *Identifying countermeasures: More important today than ever.* Presentation to the Annual Seminar of the American Polygraph Association, Reno, NV.

Department of Defense Polygraph Institute Research Division Staff (1997). A comparison of psychophysiological detection of deception accuracy rates obtained using the Counterintelligence Scope Polygraph (CSP) and the Test for Espionage and Sabotage (TES) question formats. *Polygraph*, 26 (2), 79-106.

Department of Defense Polygraph Institute Research Division Staff (1998). Psychophysiological detection of deception accuracy rates obtained using the Test for Espionage and Sabotage (TES). *Polygraph*, 27 (1), 68-73.

Fuse, L.S. (1982). *Directed lie control testing technique*. Unpublished manuscript.

Golden, R.I. (1969). The "yes-no" technique. Paper presented at the American Polygraph Association Seminar in Houston, TX.

Honts, C.R., & Raskin, D.C. (1988). A field study of the directed-lie control question. *Journal of Police Science and Administration,* 16*, 56-61.*

Honts, C.R. (1999). The discussion of questions between list repetitions (charts) is associated with increased test accuracy. *Polygraph*, 28 (2), 117-123.

Honts, C.R., & Gordon, A. (1998). A critical analysis of Matte's analysis of the directed lie. *Polygraph*, 27 (4), 241-252.

Honts, C.R., Raskin, D.C., Amato, S.L., Gordon, A., & Devitt, M. (2000). The hybrid directed-lie test, the overemphasized comparison question, chimeras and other inventions: A rejoinder to Abrams (1999). *Polygraph*, 29 (2), 156-168.

Honts, C.R., Raskin, D.C., & Kircher, J.C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology, 79*, 252-259.

Horowitz, S.W., Kircher, J.C., Honts, C.R., & Raskin, D. C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology,* 34, 108-115.

Kircher, J. C., Packard, T., Bell, B.G., & Bernhardt, P. C. (2001). Effects of prior demonstrations of polygraph accuracy on outcomes of probable-lie and directed-lie polygraph tests. DoDPI02-R-0002. DTIC AD Number A404128. University of Utah.

Krapohl, D.J., Stern, B.A., & Bronkema, Y. (2003). Numerical evaluation and wise decisions. *Polygraph,* 32 (1), 2-14.

Krapohl, D.J., & Stern, B.A. (2003). Principles of multiple-issue polygraph screening: A model for applicant, post-conviction offender, and counterintelligence testing. *Polygraph,* 32 (4), 201-210.

Lykken, David T. (1998). *A tremor in the blood: Uses and abuses of the lie detector.* Second Edition. New York: Plenum Trade.

Matte, J.A., (1998). An analysis of the psychodynamics of the directed-lie control question in the control question technique. *Polygraph*, 27 (1), 56-67.

Matte, J.A., & Reuss, R.M. (1999). Validation of potential response elements in the directed-lie control question. *Polygraph*, 28 (2), 124-142.

Matte, J.A. (2000). A critical analysis of Honts' study: The discussion (stimulation) of comparison questions. *Polygraph*, 29 (2), 146-148.

Reali, S.F. (1978). Reali's positive control technique. *Polygraph*, 7 (4), 281-285.

Reed, S. (1994). A new psychophysiological detection of deception examination for security screening. *Psychophysiology*, 31, S80. (Abstract)

Reid, J.E., & Inbau, F.E. (1977). *Truth and deception: The polygraph ("lie detector") technique.* Second edition. Baltimore, Md: Williams & Wilkins Co.

The Relationship between Facial Skin Surface Temperature Reactivity and Traditional Polygraph Measures Used in the Psychophysiological Detection of Deception: A Preliminary Investigation[1]

Dean A. Pollina, Ph.D. and Andrew H. Ryan, Ph.D.

Abstract

This study investigated the feasibility of combining traditional polygraph measures including blood volume, respiration, and electrodermal activity with facial skin surface temperature (SST) changes recorded using high definition thermal imaging. Participants were randomly assigned to nondeceptive (n = 13) or deceptive (n = 12) treatment groups using a mock-crime scenario. The frequencies of accurate determinations made using traditional polygraph measures, SST measures, and a combination of polygraph and SST measures were compared using binary logistic regression. Highest accuracy was obtained using a combination of polygraph and SST measures, suggesting that recordings of facial SST provide information that may be useful when combined with traditional measures during a polygraph examination. These results are discussed in relation to orienting response (OR) theory.

Introduction

Recent research suggests that many emotions are accompanied by specific physiological, biochemical, and behavioral responses (Ekman, Hager, & Friesen, 1981). According to Ekman, (1992) facial expressions are the richest source of information about the underlying emotional states that accompany them. There is evidence that the involuntary facial expressions of emotion are the product of evolution, and many human facial expressions are seen on the faces of other primates. Additionally, researchers have met with some success using the patterns of facial muscle activity that are ultimately responsible for the expression of an emotion to discover additional information about emotions in humans. The muscles involved in facial expression have been used in an attempt to determine which specific emotion is felt, the strength of the felt emotion, and whether more than one emotion is being experienced at a given time (Ekman, 1992; Ekman, Freiesen, & Ancoli, 1980; Sackeim, Gur, & Saucy, 1978). Importantly, this research shows that certain facial expressions of emotions are not under voluntary control and can occur even when research participants do not want them to.

The findings that certain facial expressions are involuntary and products of evolution have led several researchers to investigate the possibility that these facial expressions might betray the deceptive utterances of lairs. If so, behavioral studies of changes in facial expression might be a useful addition to the battery of psychophysiological detection of deception (PDD) measures employed by polygraph examiners (who typically monitor cardiovascular, electrodermal, and respiratory activity) as a series of questions that relate to an incident under investigation are being asked. Unfortunately, this line of research is made more difficult to interpret because involuntary facial expressions are very quickly masked. In the end, the muscles of the face are largely under voluntary control. Several studies with chimpanzees (Woodruff & Premack, 1979) and humans (Ekman & O'Sullivan, 1991) show that research participants can suppress some

[1] This work was originally report number DoDPI02-P-0007, Department of Defense Polygraph Institute, Fort Jackson, SC 29207.

of the behaviors that convey accurate information when they are attempting to deceive someone. One technology that shows promise in overcoming some of the limitations of behavioral studies of facial expression is thermography. Thermography is a technique used for measuring radiant energy or natural heat (infrared) emission from the human body (Gorbach, 1993). Using infrared (IR) radiometry, heat measurements from large areas of the body surface can be made without skin contact. Skin surface temperature (SST) is affected by changes in underlying muscle activity and microcirculation. Increases in muscle activity result in increases in blood flow to the arterioles surrounding the muscle, and these changes are associated with a rise in heat production (Grayson, 1990).

In addition to the SST changes caused by facial muscle activity, sympathetic vasodilatation and constriction are involved in facial flushing in response to body heating and embarrassment (Drummond & Lance, 1987). Fox et al., (1962) showed that vasomotor control of certain areas of the face involves variations in vasoconstriction, whereas vasomotor control of other facial areas involves vasodilatation. Both mechanisms appear to play a part in regulating the circulation of blood in the skin overlying the nose. Cervical sympathetic fibers are most likely involved in producing this vasodilatation and vasoconstriction. The distribution of heat radiating from the face is also affected by facial sweating, which, like flushing, is controlled by the sympathetic branch of the ANS. Given all the factors that can contribute to changes in facial SST, it is likely that the patterns of SST responses to questions during a polygraph examination are extremely complex. By combining the traditional ANS measures used in the psychophysiological detection of deception with rapid-response facial SST measures, it was predicted that a more complete picture of the psychophysiological processes that accompany deliberate deception would emerge. Of particular interest was the hypothesis that specific facial areas would be differentially affected by participants' fear-induced central and ANS responses to specific test questions. It was also hoped that the SST information would contribute to more effective discrimination of deceptive and nondeceptive individuals than the traditional polygraph measures alone.

Method

Participants

Thirty participants (20% Female) between the ages of 19 and 28 (M = 21.2) were recruited from a sample of U.S. Army basic trainees stationed at Fort Jackson, South Carolina and assigned to duty at the Department of Defense Polygraph Institute (DoDPI). All participants were given the option of participating in this research study, or watching television or reading in the DoDPI library for the day. Only interested volunteers were selected for participation. All participants were in good health by self-report, and no one was taking any medications except for pain killers (e.g. ibuprofen) because of minor injuries sustained during basic training. Five participants were dropped from the study due to incriminating statements made to the polygraph examiner (n = 1), sleeping during the polygraph examination (n = 2), or failure or unwillingness to commit the mock crime (n = 2). This resulted in the inclusion of 25 participants (12 deceptive, 13 nondeceptive) in the final data analyses[2].

Apparatus

Polygraph recordings were obtained using Axciton model field polygraphs (Axciton Systems, Inc, Houston TX). All examinations included measures of respiration, relative blood volume, and electrodermal activity. A Raytheon model FPA thermal imaging radiometer was used to monitor SST. The radiometer 12-bit digital output was connected to a high-speed digital video processing board chair was also used. Physiological data were collected in a darkened, temperature supplied with software designed specifically for thermal imaging installed in a Pentium III 466 Mhz computer. A mock-crime room including a plastic dummy, purse, screwdriver, and controlled room (Range: 20-22^0 C).

[2] Data on examiner decisions (deceptive / nondeceptive) for a subset of the data reported here was published previously in the journal *Nature*. The *Nature* article also reported the first use of a classification algorithm utilizing thermal imaging data in the periorbital region. See references section for a complete citation.

General Procedure

All interested participants were instructed to read a brief description of this research project and sign an informed consent form. Each participant then answered a series of biographical and medical questions to ensure that they were in good health and not taking medication that could potentially interfere with the examination results. After all forms were completed, the investigator explained how the polygraph examination would be conducted. Each participant was randomly assigned to either the deceptive or non-deceptive group. Participants in the nondeceptive group were told that they would be taking a polygraph test as part of a research study and questioned about the murder of a woman that took place at the DoDPI earlier that day. Since they did not commit this crime, they were instructed to answer all questions truthfully during the polygraph examination. Participants assigned to the deceptive group were told that they would be involved in a pretend crime, and would lie about this during the polygraph examination in an attempt to appear innocent. Participants then either waited quietly to be brought to the polygraph examination room (non-deceptive group) or committed a pretend crime (deceptive group).

Procedure for Deceptive Group

Prior to each participant's arrival at the DoDPI, a mock crime room was constructed. In the room, a plastic dummy was seated in a chair. A purse containing $20 USD was placed next to the dummy, and a screwdriver was placed on a table next to the purse. Participants in the deceptive group were instructed by the investigator to enter the mock-crime room without being seen, stab the dummy with the screwdriver, and steal the $20 from the purse. After committing the mock crime, each participant was asked details about the crime by the experimenter. Questions included, "Were you seen by anyone? Did you remember to steal the $20? What happened to the woman in the room?" Participants who failed to stab the dummy and steal the $20 were excluded from this study.

Data Collection Procedures

At a prearranged time, each participant was met by a U.S. Government certified polygraph examiner, who was blind to the participant's group membership. Sensors were attached to the participant in the following locations: electrodermal finger plates on the distal-medial phalanges of the first and third fingers of the (typically) nondominant hand, blood pressure cuff on the (typically) dominant arm above the brachial artery, and pneumographic chest assemblies across the pectoralis major muscles ("upper" pneumograph) sensor under the arm and across the rectus abdominis immediately above the navel ("lower" pneumograph) sensor. Each participant was questioned briefly about the crime, and then the polygraph test questions, in a Zone Comparison Test format (DoDPI, 1994) were reviewed. The questions asked during data collection included crime relevant questions, comparison questions, and crime-irrelevant questions (Table 1). Ten questions were presented during a single series, and each series was repeated three times during the polygraph examination. Each question was presented approximately 25 seconds after the onset of the previous question.

Data Reduction: Polygraph Measures

The upper and lower pneumograph, electrodermal, and cardiovascular responses to each question were sampled at a rate of 15 samples/s and interpolated to a rate of 60 samples/s for all subsequent analyses. Digitized (ASCII) data collected from each of the four polygraph channels during each repetition of a question sequence were standardized using z-score transformations. The standardized data were then separated by Question Type (R5, R7, R10, C4, C6, C9) using the onset of each of the examiner's questions as the beginning of the analysis interval and the onset of the examiner's next question as the end of the analysis interval. Dependent measures included maximum amplitude of the blood volume and electrodermal responses, and Euclidean distance between successive timepoints ("line length") in the upper and lower pneumograph responses (Kircher, 1983; 1984; Kircher & Raskin, 1988). Blood volume, electrodermal, and line length responses were calculated for each of the three relevant (R1, R2, R3) and comparison (C1, C2, C3) questions asked during each of the three

Table 1. *Questions Asked by the Polygraph Examiner*

Test Question	Question Type
Is your Name _____?	Not Scored
Regarding whether you stabbed that woman today, do you intend to answer my questions about that truthfully?	Not Scored
Do you understand that I will not ask any trick or surprise questions on this test?	Not Scored
Before arriving at Fort Jackson, did you ever seriously hurt someone who trusted you?	C1
Did you stab that woman today?	R1
Before arriving at Fort Jackson, Did you ever lose your temper when you shouldn't have?	C2
Did you stab that woman in that room today?	R2
Is there anything you are afraid I will ask you a question about even though I said I wouldn't?	Not Scored
Before this year, did you ever take anything important that didn't belong to you?	C3
Do you have that stolen $20 on you right now?	R3

question series presented during the polygraph examination.

To determine maximum amplitude of the blood volume response, difference scores in relative units were obtained between each low point and successive high points identified in each response curve using a 13.7 s analysis window. Maximum amplitude of the blood volume response was defined as the greatest such difference. Maximum amplitude of the electrodermal response was determined in a similar manner with a 5 s analysis window for determination of minimum amplitude. Length of the upper and lower pneumograph tracing was determined by measuring the Euclidean distance between successive pairs of samples obtained every 1/60 s for 10 poststimulus seconds. The resulting 600 measurements were summed to yield a length measure in relative units for each respiration channel. Difference scores were obtained for blood volume response measures by subtracting, for each subject, mean responses to each comparison question from mean responses to each adjacent relevant question. Similar subtraction measures were also derived for electrodermal, and upper and lower pneumograph responses. By collapsing across question type and question sequence yielded a single (comparison question – relevant question) subtraction measure for each participant.

Data Reduction: SST Measures

Recordings of facial temperature values were started at the onset of each of the examiner's relevant (R1, R2, R3) and comparison (C1, C2, C3) questions using a 30 Hz sampling rate for ten seconds (300 image frames) and a 256 x 256 FPA. Thermal image data collection was started with the press of a computer key after a prearranged signal (finger tap) from the polygraph examiner prior to beginning each question in the sequence. The resulting (256 x 256 x 300) array of temperature values collected during the presentation of each question was converted to an ASCII text file and stored on a CD-RW disc for off-line data analysis. Three "subtraction" temperature arrays were created for each question sequence by subtracting each point in the relevant question temperature array from the corresponding point in the adjacent comparison question temperature array. Due

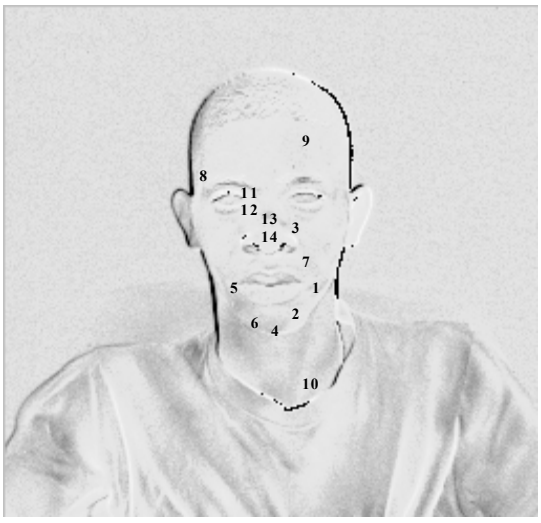
to data storage limitations, SST data were not collected during the third repetition of the question sequence. This resulted in a total of six subtraction temperature arrays generated for each participant.

To test the hypothesis that comparison question – relevant question temperature differences across the face are related to deception, temperature/timepoint waveforms from fourteen facial areas were selected for analysis. All fourteen of the selected facial areas overlie muscles involved in facial expression (Martini, 1998). To determine maximum amplitude of the SST responses at each facial area, difference scores in relative units were obtained between the lowest and highest point on each waveform using the 10 s analysis window in each subtraction array (Figure 1). For those facial areas that were bilaterally symmetric (e.g., mouth, ears, neck, eyes, scalp), SST maximum amplitudes were collected on both the left and the right side of the face. The average of each pair of measures was used in all subsequent statistical analysis. Finally, average waveforms were created by collapsing across facial areas overlying muscles controlling the mouth (Areas 1- 7), ear (Area 8), scalp (Area 9), neck (Area 10), eye (Areas 11 and 12), and nose (Areas 13 and 14).

Results

Facial SST Maximum Amplitudes. Figure 2 shows grand average waveforms, collapsing across individual participants' responses, at each of six facial areas. Reliable SST waveforms were generated to all comparison and crime-relevant questions asked during the first two repetitions of the ZCT test. Visual inspection of the grand average waveforms suggests that skin surface regions in areas around the nose and eyes manifested the largest amplitude group differences. It also appears that SST response latencies differed at each facial region. Specifically, peak latencies appear as a traveling wave, with shortest peak latencies at the ear, scalp and nose. Peak latencies are longer in the mouth and neck regions, and the periorbital eye regions showed the longest peak latencies in the grand average waveforms (Figure 2). A between-groups multivariate analysis of variance was performed on the six dependent variables associated with the thermal imaging measures. The combined thermal imaging DVs were not significantly related to (Deceptive/ Nondeceptive) group membership $F(6, 17) = .29$, n. s, $\eta^2 = .09$.

Figure 1. First frame in (comparison question – relevant question) subtraction SST array which began at the onset of a single question spoken by the polygraph examiner. Numbers indicate sites chosen for thermal image analysis. All sites were skin surface areas overlying muscles involved in facial expression.



Mouth
1. Buccinator.
2. Depressor labii inferioris.
3. Levator labii superioris.
4. Mentalis
5. Orbicularis oris
6. Depressor anguli oris.
7. Zygomaticus major and minor.

+ .60^{0}C
0
- .60^{0}C

Ear
8. Temporoparietalis.

Scalp
9. Frontalis.

Neck
10. Platysma.

Eye
11. Corrugator supercilii
12. Orbicularis oculi

Nose
13. Procerus.
14. Nasalis.

Figure 2. Grand Average SST waveforms, collapsing across individual participants' responses, at each of six facial skin surface areas, including mouth, ear, neck, nose, eye, and scalp. Temperature changes appear as a traveling wave from time intervals 1 – 6.

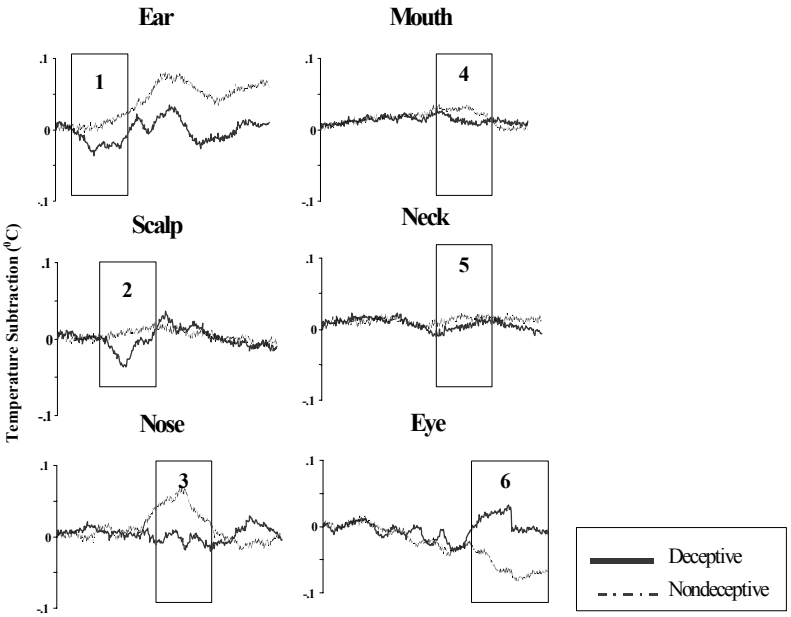


Table 2. *Mean (+/- S. E. M.) SST and Polygraph Subtraction Measures*

Comparison – Relevant Subtraction Measure	Deceptive Participants	Nondeceptive Participants
Maximum SST Amplitude*		
Mouth	20.96 (+/- 1.68)	18.99 (+/- 1.52)
Ear	23.66 (+/- 4.87)	21.86 (+/- 2.87)
Scalp	17.00 (+/- 3.94)	14.08 (+/- 1.47)
Neck	12.64 (+/- 1.62)	11.02 (+/- 1.03)
Eye	39.64 (+/- 5.44)	43.42 (+/- 7.60)
Nose	22.35 (+/- 3.79)	24.05 (+/- 4.74)
Polygraph Measures		
Maximum EDA Amplitude	-.83 (+/- .15)	.07 (+/- .23)
Maximum BV Amplitude	-.90 (+/- .35)	-.66 (+/- .36)
Respiration Line Length (Upper)	.37 (+/- .30)	.04 (+/- .26)
Respiration Line Length (Lower)	.49 (+/- .34)	.05 (+/- .32)

Note. The units of measure for thermal data are digital values recorded from the camera. An increase of one digital unit is ≈ .013 degree Celsius.

Polygraph Measures. Maximum amplitude of the blood volume and electrodermal responses, as well as respiration line length all showed the predicted response patterns. Specifically, maximum amplitudes of the blood volume and electrodermal responses were larger to the relevant questions in deceptive participants, and larger to the comparison questions in nondeceptive participants (Table 2). Respiration line length showed the (predicted) opposite pattern. Respiration line lengths decreased to a greater degree in response to the relevant questions in deceptive participants, and decreased to a greater degree in response to the comparison questions in nondeceptive participants. A between-groups multivariate analysis of variance was performed on the four dependent variables associated with the traditional polygraph measures. With the use of Wilks' criterion, the combined polygraph DVs were significantly related to (Deceptive/ Nondeceptive) group membership $F(1, 22) = 5.61$, $p < .01$, $\eta^2 = .54$.

SST/Polygraph Classification Accuracy.

The ability to classify individuals into deceptive and nondeceptive groups based on blood volume, electrodermal, and pneumograph measures was examined using binary logistic regression. Maximum amplitude of the blood volume response, maximum amplitude of the electrodermal response, and (upper and lower) pneumograph Euclidean distance (relevant question – comparison question) subtraction measures were entered into the logistic regression equation as each of four covariates in a single block. As predicted, these four variables accounted for a significant proportion (R^2 = .41) of the variation in the regression model, X^2 (4, N = 24) = 12.6, $p < .01$. Next, maximum amplitude of the skin temperature responses from the nose, mouth, neck, eye, scalp, and ear (relevant question – comparison question) subtraction measures were entered as each of six covariates in a separate logistic regression analysis. These six SST variables were entered in a single block. These variables failed to account for a significant proportion (R^2 = .09) of the variation in the model, X^2 (6, N = 24) = 2.3, n. s.

To determine whether an interaction between traditional polygraph measures and SST amplitudes could result in better predictive value than using either approach alone, the four polygraph measures (upper and lower pneumograph line length, blood volume and electrodermal maximum amplitude) were combined with nose SST maximum amplitude in a two-step binary logistic regression analysis. Using these five measures as covariates, the logistic regression equation again accounted for a significant proportion (R^2 = .49) of the variation in the regression model, $X^2(5, N = 24) = 16.0$, $p < .01$. This logistic regression equation also reached significance when the eye SST maximum amplitude measure was substituted for nose SST at step 2, $X^2(5, N = 24) = 14.6$, $p < .01$, and again when nose SST and eye SST were both entered at step 2, $X^2(6, N = 24) = 17.5$, $p < .01$. However, the increase in the proportion of variance accounted for by the SST variables entered at step 2 failed to reach statistical significance, $X^2(2, N = 24) = 2.8$, n.s.

Next, the *c* statistic, equal to the area under the ROC curve, was calculated for each of the logistic regression analyses previously performed (See Hanley and McNeil, 1982 for an explanation of the *c* statistic). Case inclusion cutoff probability was set at .5, with probabilities < .5 classified as deceptive and probabilities > .5 classified as nondeceptive. Results show a shift in the probability distribution toward correct classification when both traditional polygraph measures and SST measures are combined (Table 3). Logistic regression analysis generates a direct estimate of the probability of an event occurring. This feature allows for an arbitrary case inclusion cutoff probability, which was included in the above analysis to generate an "inconclusive" column. Typically, a certain number of field polygraph examiners' test results are inconclusive. Table 4 shows the number of correct, incorrect, and inconclusive (no opinion) results at .50, .60, .70, .80 and .90 cutoff probability levels. Results indicate that the combination of Polygraph measures combined with SST from the eye and nose facial regions yields the most robust discrimination of group membership, especially at the more stringent criteria.

Table 3. *Area Under the ROC Curve Derived from Binary Logistic Regression*

Regression Analysis	Predictor Variables	R^2 (Cox & Snell)	ROC Area	Sig. *
Polygraph Measures	BV, EDA, AR, TR	.41	.88	.002
SST Amplitude Measures	SST: Nose, Mouth, Eye, Scalp, Neck, Ear	.09	.70	.09 (N.S.)
Polygraph and SST Amplitude: Nose	BV, EDA, AR, TR, SST: Nose	.49	.90	.001
Polygraph and SST Amplitude: Eye	BV, EDA, AR, TR, Eye	.46	.90	.001
Polygraph and SST Amplitude: Eye, Nose	BV, EDA, AR, TR, SST: Eye, Nose	.52	.92	.001

*Null hypothesis: true area = .50

The increases in classification accuracy achieved with the combination of polygraph and SST measures suggest that a degree of orthogonality exists among the polygraph and SST measures. To examine this, bivariate (Pearson) correlations were performed between a composite SST measure (collapsing across eye and nose) and cardiovascular, electrodermal, and respiration polygraph measures. Using the entire sample, none of these correlations reached significance at the .05 level. However, an additional analysis conducted on the subgroups of deceptive and nondeceptive individuals showed an interaction between maximum amplitude of the cardiovascular response to crime-relevant questions and composite SST maximum amplitudes to relevant questions. Deceptive individuals whose cardiovascular responses were least responsive to the crime relevant questions had the most responsive SST responses to crime-relevant questions. Nondeceptive individuals with largest cardiovascular responses to the crime-relevant questions also had the most responsive SST responses to the relevant questions (Figure 3).

To test whether the effect shown in Figure 3 was statistically significant, raw values from each participant's SST and Cardiovascular subtraction measures were re-coded into a single dichotomous variable. Participants with negative CQ – RQ subtraction scores and low CQ – RQ SST (eye and nose) received a value of +1. Participants with positive CQ – RQ subtraction scores and high CQ – RQ SST (eye and nose) received a value of +1. Participants with negative CQ – RQ subtraction scores and high CQ – RQ SST (eye and nose) received a value of -1. Participants with positive CQ – RQ subtraction scores and low CQ – RQ SST (eye and nose) received a value of -1. A Spearman correlation (phi-coefficient) between these re-coded values and (Deceptive/ Nondeceptive) group membership was significant (r = .43, $p < .05$). This result is shown in Table 5.

Table 4. *Classification Table: Binary Logistic Regression*

Cutoff Score	Polygraph Measures			Polygraph & SST: Eye			Polygraph & SST: Nose			Polygraph & SST: Eye, Nose		
	Hit	Miss	Inc.	Hit	Miss	Inc.	Hit	Miss	Inc.	Hit	Miss	Inc.
.5/.5	16	8	0	19	5	0	21	3	0	20	4	0
.6/.4	16	4	4	17	2	5	16	3	5	18	3	3
.7/.3	15	2	7	15	2	7	14	1	9	16	2	6
.8/.2	14	0	10	12	2	10	12	1	11	15	0	9
.9/.1	7	0	17	11	0	13	12	0	12	12	0	12

Figure 3. Scatterplot showing the relationship between blood volume measures derived from sphygmomanometer recording over brachial artery and skin surface temperature recordings from the eye and nose regions of the face. Fit-lines through each group's data were calculated using a lowess model with 50% points-to-fit.

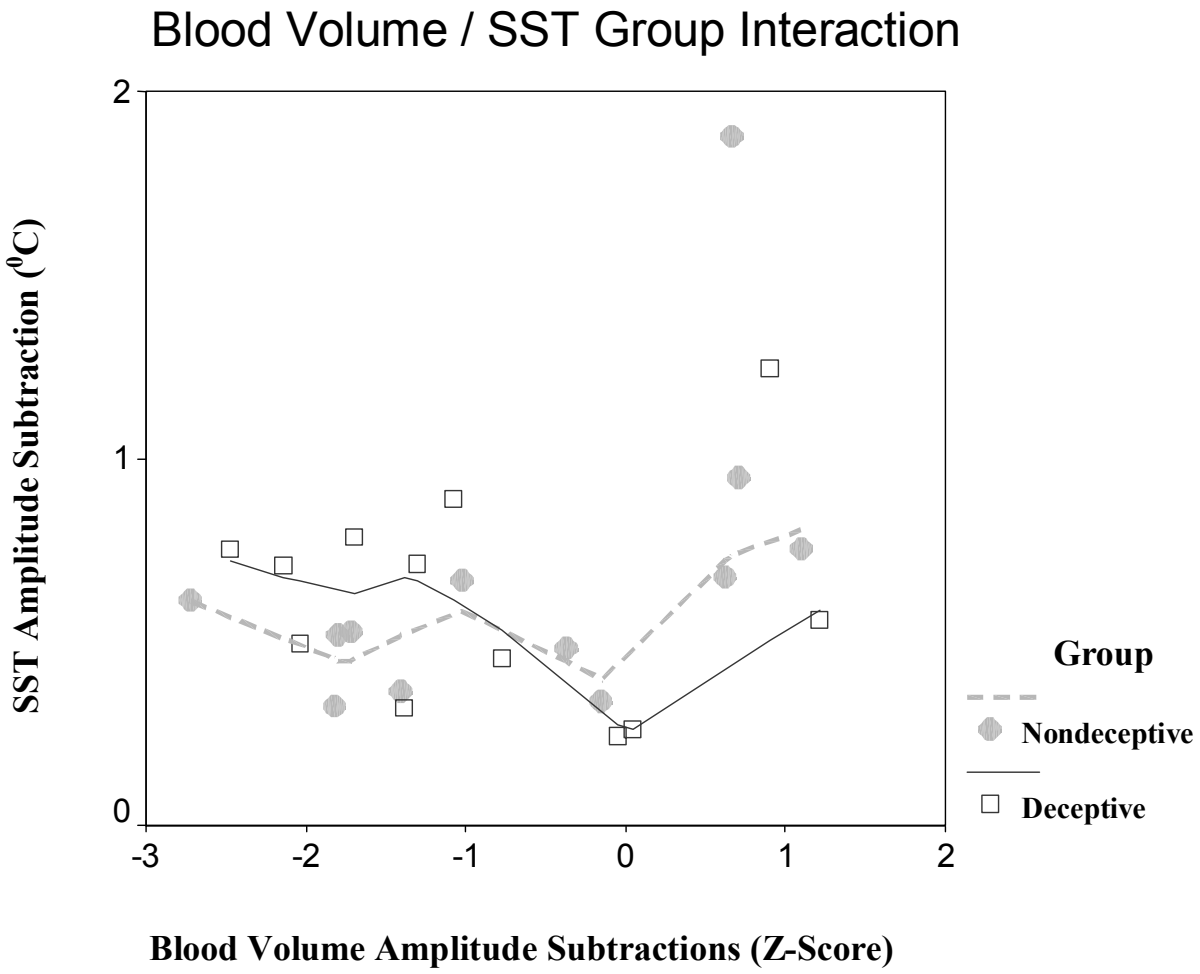


Table 5. *Combined Blood Volume and SST Measures For each Group*

Re-coded Variable	*Group Assignment*			
	Deceptive		Nondeceptive	
	Count	%	Count	%
-1	7	58.3	2	16.7
+1	5	48.7	10	83.3

Note. High CQ – RQ SST and High CQ – RQ BV = +1; Low CQ – RQ SST and Low CQ – RQ BV = +1; High CQ – RQ SST and Low CQ – RQ BV = -1; Low CQ – RQ SST and High CQ – RQ BV = -1

Discussion

The results of this study suggest that thermal image analysis, when combined with traditional polygraph measures, can be effective in discriminating deceptive and nondeceptive individuals during a polygraph test. Using high-definition, rapid response thermal imaging, real-time changes in skin surface temperature across the face were effectively recorded. Skin surface temperatures overlying muscles around the eyes and nose appeared to be the most effective predictors of deception, but only when combined with the traditional polygraph measures of respiration, cardiovascular, and electrodermal activity. The finding that dichotomous classifications based on a combination of SST and BV amplitudes significantly correlated with group membership also suggests that a combination of SST and traditional polygraph measures might lead to more effective means of detecting deception. These findings also give some support to the theory that unique facial heat signatures and perhaps neuromuscular patterns are associated with specific emotions (Ekman, Hagar, & Friesen, 1981; Ekman, Levenson, and Friesen, 1983; Pavlidis, Eberhardt, & Levine, 2002). However, more work will be necessary to effectively determine the relationship between cardiovascular and neuro-muscular activity, SST, and emotional states.

Using measures of vagal tone derived from the electrocardiogram, Raskin and Kircher (1990) found that deceptive participants showed less vagal response to relevant questions than to comparison questions. Nondeceptive participants showed the opposite pattern. Vagal tone is a measure of parasympathetic (vagus nerve) influence on heart rate (Porges et al., 1980). However, a subsequent study failed to find any significant correlation between vagal tone measures and ground truth using a mock crime scenario (Kircher, Packard, Bell, & Bernhardt, 2001). These discrepant findings suggest that the contribution made by the parasympathetic branch of the ANS to the psychophysiological detection of deception is modest when standard cardiovascular measures are used. However, measures such as SST, which can be taken from many regions of the body simultaneously, could be effective in revealing the differential effects of the sympathetic and parasympathetic branches of the ANS on microcirculation in capillaries near the skin's surface. One possibility is that the SST/BV interactions seen in the present study are the result of combined sympathetic and parasympathetic effects on SST. Future studies investigating the separate component processes that underlie facial SST during a PDD examination could be useful in answering this question.

Several fundamental emotions have been shown to exist across different cultures, and attempts have been made to link these emotions to specific facial expressions (Ekman, Friesen, & Ellsworth, 1972; Izard, 1972). The face has a high ratio of motor units to muscle mass and extensive neural innervation, and facial EMG patterning has been shown to be sensitive to different emotional states elicited by affective imagery (Schwartz, Fair, Salt, Mandel, & Klerman, 1976). These patterns are not typically noticeable in the overt face (Schwartz, 1986). The results of the present study suggest that SST patterns might also be sensitive to specific emotions, but the exact nature of the resulting temperature patterns is not yet known. Sites chosen for SST analysis in the present study were those that overlie muscles involved in facial expression. However, it is still unclear whether the observed temperature changes are highly correlated with activity in specific muscle groups. The two sites most predictive of deception in the present study were the periorbital regions around the eyes and the nose.

Although there is no unified theory that explains the effectiveness of the PDD process, most parsimonious explanations involve orienting and defensive responses (Sokolov, 1963, Sokolov & Cacioppo, 1997). Changes in blood volume are part of the orienting response (OR) first described by Sokolov, (1963), who reported decreases in forehead blood volume in response to threatening stimuli. According to OR the theory proposed by Sokolov, these decreases in cephalic blood volume reflect a defensive response (DR) that protects the organism from harm. Conversely, novel or unexpected stimuli produce increases in forehead blood volume which reflect an

orienting response that improves perceptual ability. More recent studies investigating orienting and defensive responses have shown that individual differences play a role in cardiac responses to fear stimuli. In one study, subgroups of individuals showed acceleratory, deceleratory, and moderate deceleratory responses to pictures of homicide victims (Hare, 1972). These results have been interpreted to suggest that a given stimulus can evoke DRs in some participants, and ORs in other participants (Cook & Turpin, 1997). The results of the present study also suggest that, in the forehead and periorbital region, the situation is complex. A multivariate approach to the study of facial SST, based on the principle of a multidimensional space including sympathetic-parasympathetic inputs to the heart and face may be useful in determining the components of the observed facial temperature distribution in response to threatening stimuli.

Acknowledgments

The authors would like to thank Kay Williams and Rose Swinford of the DoDPI Research Staff for their assistance with data collection procedures. We would also like to thank Gordon Barland and Don Krapohl, both of whom administered the polygraph exams to our study participants. We are also grateful to Johnnie Rodgerson for his advice concerning instructions given to deceptive participants, and Peter Reutiman for his technical assistance. This project was funded by the Department of Defense Polygraph Institute as project number DoDPI00-P-0011. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Cook, E. and Turpin, G. (1997). Differentiating orienting, startle, and defense responses: The role of affect and its implications for psychopathology. In: P. Lang, R. Simons, & M. Balaban, (Eds.), *Attention and Orienting: Sensory and Motivational Processes*. New Jersey: Lawrence Erlbaum Associates.

Department of Defense Polygraph Institute. (1994). *ZCT, Zone Comparison Test*. Unpublished Manuscript, Department of Defense Polygraph Institute.

Drummond, P., and Lance, J. (1987). Facial flushing and sweating mediated by the sympathetic nervous system. *Brain*, *110*, 793-803.

Ekman, P. (1992). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton & Company.

Ekman, P., Friesen, W., and Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*, 1125-1134.

Ekman, P., Friesen, W., and Ellsworth, P. (1972). *Emotion in the Human Face*. Elmsford NY: Pergamon Press.

Ekman, P., Hager, J., and Friesen, W. (1981). The symmetry of emotional and deliberate facial actions, *Psychophysiology*, *18*, 101-106.

Ekman, P., Levenson, R. & Friesen, W. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *22*, 1208-1210.

Ekman, P., and O'Sullivan, M. (1991). Who can catch a liar. *American Psychologist*, *46*, 913-920.

Fox, R., Goldsmith, R., and Kidd, D. (1962). Cutaneous vasomotor control in the human head, neck, and upper chest. *Journal of Physiology, 161*, 298-312.

Grayson, J. (1990). Responses of the microciculation to hot and cold environments. In: W.C. Bowman, E. Schönbaum, & P. Lomax (Eds.), *Thermoregulation: Physiology and Biochemistry* (pp. 221-234). New York: Pergamon Press.

Gorbach, A. M. (1993). Infrared imaging of brain function. In: U. Dirnagl, A. Villringer, & K.M. Einhaupl (Eds.), *Optical Imaging of Brain Function and Metabolism* (pp. 95-123). New York: Plenum Press.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver-operating characteristic (ROC) curve. *Radiology, 143*: 29-36.

Hare, R. (1972). Cardiovascular components of orienting and defensive responses. *Psychophysiology, 9*, 606-614.

Izard, C. (1972). *Patterns of Emotions*. New York: Academic Press.

Kircher, J. C. (1983). *Computerized decision-making and patterns of activation in the detection of deception*. Unpublished doctoral dissertation, University of Utah.

Kircher, J. C. (1984). Uses and abuses of the mock crime paradigm in research on field polygraph techniques. *Psychophysiology, 21*, 566.

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.

Kircher, J. C., Packard, T., 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. DoDPI-02-R0002). *University of Utah*, Department of Educational Psychology.

Martini, F. H. (1998). *Fundamentals of Anatomy and Physiology* (4th Ed.). New Jersey: Prentice Hall, Inc.

Pavlidis, I., Eberhardt, N. L., & Levine, J., (2002). Human behavior: Seeing through the face of deception. *Nature, 415*, 35.

Porges, S., Bohrer, R., Cheung, M., Drasgow, F., McCabe, P., & Keren, g. (1980). New time series statistic for detecting rythmic co-occurrence in the frequency domain: The weighted coherence and its application to psychophysiological research. *Psychological Bulletin, 88*, 580-587.

Raskin, D., & Kircher, J. (1990). Development of a computerized polygraph system and physiological measures for detection of deception and countermeasures: A pilot study (Contract 88-L55300-000). Salt Lake City: *Scientific Assessment Technologies*.

Sackeim, H., Gur, R., and Saucy, M. (1978). Emotions are expressed more intensely on the left side of the face, *Science, 202*, 434.

Schwartz, G. (1986). Emotion and psychophysiological organization: A systems approach. In M. Coles, E. Donchin, & S. Porges, (Eds.) *Psychophysiology: Systems, Processes, and Applications*. New York: Guilford Press.

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., & Klerman, G. L. (1976). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science*, *192*, 489-491.

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

Sokolov, E. and Cacioppo, J. (1997). Orienting and defense reflexes: vector coding the cardiac response. In: P. Lang, R. Simons, & M. Balaban, (Eds.), *Attention and Orienting: Sensory and Motivational Processes*. New Jersey: Lawrence Erlbaum Associates.

Woodruff, G., and Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, 7, 333-362.

When Did You Conclude She was Lying? The Impact of the Moment the Decision about the Sender's Veracity is Made and the Sender's Facial Appearance on Police Officers' Credibility Judgments[1]

Jaume Masip, Eugenio Garrido[2], and Carmen Herrero

Abstract

Two experiments were conducted to explore how the moment observers make their decision about the senders' veracity affects their judgment and detection accuracy. In Experiment 1 police officers and undergraduates judged the credibility of video-recorded statements. Contrary to our expectation, officers did not judge the statements earlier than the students. An initial lie bias became evident. In Experiment 2 a still face, which could be of the same witness as in Experiment 1, or of two other witnesses, was shown to officers as they listened to truthful or deceptive accounts taken from Experiment 1. There was no effect of the sender's facial appearance on the lie bias found in the first experiment, which emerged here as well. Accuracy for detecting deceptive accounts decreased across time in both studies, while accuracy for truthful accounts increased only in Experiment 2. How visual and verbal information contributed to these effects is discussed.

Introduction

DePaulo and Rosenthal (1979) identified three main areas of inquiry in the field of nonverbal detection of deception: (a) people's ability to lie successfully and to accurately detect deception, (b) channel or modality effects on accuracy, i.e., what kind of information (visual, vocal, verbal, transmitted by the face, transmitted by the body, etc.) is most useful for untrained observers to detect deception, and (c) the study of the behavioral indicators of deception (real deception cues, perceived deception cues, and behaviors believed by people to be useful to detect deception). Within the first area pointed out by DePaulo and Rosenthal (1979), attention has been paid to sender and/or receiver variables that may affect their ability to deceive or detect deception (variables such as gender, age, experience, personality traits, etc.), as well as to certain situational variables such as motivation to lie successfully, familiarity

[1] This article is reprinted with permission from the Journal of Credibility Assessment and Witness Psychology taken from Volume 4, Number 1, pages 1 - 36. Please refer to http://truth.boisestate.edu for copies for reproduction under the original copyright that follows:

[2]Correspondence concerning this article should be sent to Jaume Masip (jmasip@usal.es) or Eugenio Garrido (garrido@usal.es), Department of Social Psychology and Anthropology, University of Salamanca, Facultad de Psicología, Avda. de la Merced, 109-131, 37005 Salamanca (Spain).

The research reported here was supported by the Junta de Castilla y León, Programa de Apoyo a Proyectos de Investigación, Ref. 30/98
.

between sender and receiver, time to create a deceptive story, perceived consequences of being detected, etc. (see reviews by DePaulo, DePaulo, Tang, & Swaim, 1989; DePaulo, Stone, & Lassiter, 1985; Ekman, 1992; Ekman & O'Sullivan, 1989; Ford, 1996; Kalbfleisch, 1992; Köhnken, 1989; Kraut, 1980; Masip & Garrido, 2000, 2001a; Miller & Burgoon, 1982; Miller & Stiff, 1992, 1993; Vrij, 1998, 2000; Zuckerman, DePaulo, & Rosenthal, 1981; Zuckerman & Driver, 1985). In general, meta-analyses on the results obtained from this approach show that detection accuracy (i.e., accuracy at detecting both truths and lies, Miller & Stiff, 1993) by untrained detectors usually falls between 45 % and 60 % correct classifications, where 50 % is the chance level. In addition, it has been found that police officers are no more accurate than lay people in their credibility judgments (e.g., DePaulo & Pfeiffer, 1986; Ekman & O'Sullivan, 1991; Garrido, Masip, Herrero, & Tabernero, 1997; Garrido, Masip, & Herrero, 2003; Henderson & Hess, 1982; Köhnken, 1987; Kraut & Poe, 1980; Sanderson, 1978, cited by Bull, 1989; Vrij, 1992; Vrij & Graham, 1997; see reviews by Bull, 1989, Garrido & Masip, 1999, and Vrij, 2000). Instead, there is some evidence that police officers may even be less precise than non-officers, due to a lie bias they may display when making their judgments (Garrido et al., 1997; Sanderson, 1978, cited by Bull, 1989).

In view of that poor accuracy level among observers trying to discern whether someone is lying or telling the truth, Miller and Stiff (1993) suggested that that issue should be considered from an alternative perspective: instead of investigating detection accuracy, researchers should identify explanations for observers' errors in their credibility judgments. In line with that suggestion, in the two experiments reported here we try to identify some factors that may have an effect upon observers' accuracy at judging credibility; specifically we are trying to discern what processes underlie the poor performance attained by police officers in Garrido et al.'s (1997, 2003; see also Masip, 2002) study. In that experiment, officers' detection accuracy did not differ significantly from chance level, while students' accuracy was significantly above chance. The poor accuracy among officers was due to their tendency to judge all statements as false. Officers' accuracy at judging deceptive statements was as high as that of undergraduates, while their accuracy at judging honest statements was poorer. In fact, the tendency of police officers to judge statements as deceptive was the same regardless of the real quality (truthful or deceptive) of those statements, while students were somewhat more sensitive to the real truth value of the stories.

In an attempt to have a closer look at that lie bias among officers, Garrido and Masip (2001) explored whether it was due to a reduced capacity among them to perceive a general expressive pattern as defined by Becerra, Sánchez, and Carrera (1989). These authors suggested that the accurate detection of deceit is based on observers' perception of a general expressive pattern in the sender's behavior, a pattern that changes as the statement quality (value of truth) varies. If so, it could be the case that officers did not perceive that pattern. There may be several reasons for that. For instance, police officers may have a stronger Generalized Communicative Suspicion (GCS) than non-officers. Levine and McCornack (1991) differentiated between GCS and situationally-aroused suspicion or "state" suspicion. The former would be a "predisposition toward believing that the messages produced by others are deceptive" (Levine & McCornack, 1991, p. 328), and is described as a relatively enduring and cross situational cognitive construct. On the other hand, situationally-aroused or state suspicion is prompted by certain contextual cues. It was defined by Levine and McCornack (1991) as "a belief that communication within a specific setting and at a particular time may be deceptive" (p. 328). Unlike GCS, state suspicion is transitory and is based upon certain situational variables.

Detecting deceit is an important task for police officers. During their daily work, they are often involved in social interactions where mistrust and lack of confidence are normal, and where they must question the interviewee's assertions. That is to say, situations where a state suspicion is aroused. Yet, this suspicion, given its frequency in police work, could become chronic, arousing among officers a belief that the interviewee is

probably not being truthful. This process would end up generating a kind of suspicion that would no longer be a response to contextual cues nor would it be transitory anymore. Rather, that suspicion would be a GCS. Research has shown that high GCS ratings are associated with a tendency to make judgments of deceptiveness (Levine & McCornack, 1991). With regard to our police officers, it could be the case that their generalized suspicion prevented them from scrutinizing the witness's behavioral displays, thus not being able to perceive his or her general expressive pattern. If this were actually the case, then perhaps officers made only a biased "guess" based on their initial suspicion. Alternatively, police officers' generalized suspicion may have given rise to a confirmation bias, making them attentive to only those behaviors supporting their view that the sender was lying. In either case, police officers would be unable to perceive the general pattern described by Becerra et al. (1989). However, Garrido and Masip's (2001) results showed that not only officers, but also non-officers, were unable to perceive any general expressive pattern in the sender's behavior. Thus, that factor cannot account for the differences between the police and lay people.

In this paper we describe some further explorations of the processes underlying officers' lie bias in Garrido et al.'s (1997, 2003) study. Experiment 1 looks at whether police officers and students came to their conclusion about the sender's veracity at different times, and whether this can account for the judgmental differences between these groups which were detected by Garrido and his colleagues. Also, an interesting question is how the moment observers come to a conclusion about whether the sender is lying or telling the truth affects detection accuracy, that is, is there any point in time where accuracy is higher? Experiment 2 is a follow-up study to answer some questions raised by the results obtained in Experiment 1. Thus, the contribution of the sender's facial appearance and that of her dynamic nonverbal behavior to the profile found in Experiment 1 is explored.

Experiment 1

The availability of information (both useful and misleading behavioral cues) depends upon the moment observers make their decision about the truthfulness of the senders' account. If receivers decide at the very beginning of a sender's performance, the amount of available verbal and nonverbal information from that sender will be very limited. Conversely, if observers decide after the sender's performance has concluded, they will be able to take into account all the verbal and nonverbal behavior displayed by that sender throughout his or her performance. Thus, if information gathered by observers paying attention to the senders' behavior is used as a basis for making veracity judgments, accuracy will probably be influenced by the moment observers conclude that the sender is lying or telling the truth, at the beginning, middle, or end of his or her performance.

A first interesting question is whether police officers and non-officers (undergraduate psychology students) tend to decide at different moments in time. Such a difference might account for the differences between those groups found by Garrido et al. (1997, 2003). Our prediction concerning the moment variable is based on the contributions of Levine and McCornack when conceptualizing their GCS, as well as on the work of Stiff, Kim and Ramesh (1992). Thus, it could be the case, for instance, that officers have a strong generalized communication suspicion as mentioned above, so that they enter the situation with the a priori belief that the sender is lying, while lay observers are more attentive to the sender's behavioral displays. In that case, officers would tend to decide quickly at the beginning of the statement, because they "would be certain of it" and would see no need to pay attention to the witness's behavior, while students would tend to decide later in the sender's performance, after paying close attention to that witness's behavior and after having processed the information so gathered. Also, Stiff et al.'s (1992) paper permits drawing an alternative process, which would lead to the same prediction. Those authors justify the development and existence of a cognitive heuristic which would lead relational partners –among which mutual confidence and trust are the norm, as well as necessary to maintain

the relationship– to judge the other member's performance as truthful, without even processing the information conveyed by that other member which could potentially be relevant to judge his or her credibility. Probably the same rationale could be used to account for officers' judgmental tendencies, but in the opposite: instead of those cooperative interactions characterized by relational intimacy and trust that relational partners are involved in, police officers often get into interactions where distrust and suspicion are usual. This could create among officers a belief that the interviewee is not being truthful, in the same way that a belief that the other person is being truthful is aroused among relational partners. And, in the same way that a potential lie detector involved in a close relationship bases his or her credibility judgments on the a priori belief that his or her partner is honest, thus making heuristic judgments of truthfulness without even processing the incoming information, police officers could do something similar to conclude that the witness is being deceptive. On the other hand, lay observers judging the credibility of strangers' statements would use a rather different strategy. They would be less biased than officers concerning the sender's honesty, they would be less confident than officers in their skills to assess other people's credibility (Garrido et al., 2003), and, therefore, they would be more willing than officers to attend to and to take into account the behaviors displayed by the witness during his or her statement.

Both processes, the one based on a GCS among officers and the one derived from Stiff et al.'s (1992) findings concerning lie detection amongst relational partners, suggest that police officers will display a lie bias such as that found by Garrido et al. (1997, 2003) while lay observers will be able to take into account information drawn from the sender's behavior to make their judgment, which will make them more accurate at assessing credibility. Thus, our first hypothesis predicts that police officers will make their decision about the sender's veracity earlier than non-officers, because, unlike these, officers will tend not to pay attention to the incoming information which would help them make an accurate judgment.

An important moderating factor on the effects of the moment observers decide on judgmental accuracy may be the value of truth of the statement to be judged. For instance, it could be the case that, for truthful accounts, deciding at the end is beneficial, given the greater amount of accurate information available at that later point. However, the prediction is different when it comes to assessing deceptive statements. Liars may monitor their behavior in order to give a plausible false account (information management) as well as an honest impression (image management and behavior management) (see Buller & Burgoon, 1996, 1998; DePaulo, 1991, 1992; DePaulo & Kirkendol, 1989; Greene, O'Hair, Cody, & Yen, 1985; Masip & Garrido, 1999, 2000, 2001a; Vrij, 1998, 2000; Zuckerman et al., 1981). Therefore, the later observers come to a conclusion about the deceiver's truthfulness, the more misleading cues that deceiver will have had a chance to display. Not all of his or her cues will be misleading, since some behaviors are hardly controllable (e.g., Ekman, 1992; Ekman & Friesen, 1969, 1974), but in any case, later in the sender's performance, the amount of misleading information will be greater in false accounts than in honest ones, while the amount of truthful information will be relatively smaller. However, earlier in the account these differences will be less pronounced. Therefore, an interaction between the moment observers decide and the value of truth of the statements could be expected. Thus, our second hypothesis predicts that, as observers decide later in time, there will be a relative increase in accuracy at detecting truthful accounts and a relative decrease in accuracy at detecting deceptive accounts.

It is important to stress that this study was designed to test hypothesis one. Since only one sender was used, either supportive or non-confirmatory evidence for hypothesis two must be taken only as preliminary and suggestive evidence until replications with a large number of senders be conducted.

Method

Participants

The sender was a female undergraduate student of psychology at a Spanish University. Observers were 121 police

officers studying to become police inspectors at the Police Academy of Ávila (Spain), and 147 undergraduate students of psychology at a Spanish University[3].

Procedure

In order to increase the ecological validity of this study, we addressed some of the concerns expressed by various authors in this area (e.g., Köhnken, 1987, 1989; Miller & Stiff, 1993) by (a) motivating our senders to be convincing, (b) making the content of the statements relevant to police interrogation settings: the topic was the reporting of criminal actions (factual descriptions), (c) by giving senders a few minutes to prepare before giving their statements, (d) by having observers make a dichotomous decision ("true" or "false") instead of rating the degree of truthfulness or deceptiveness, and (e) by showing observers only two statements of some length (no less than two minutes). Normally, in laboratory research on nonverbal detection of deception a large number of small behavioral samples are shown to observers. However, in the real world officers rarely have to judge the credibility of dozens of statements that are only a few seconds long. We addressed this issue by showing observers only two statements of some length, although this prevented us from using a large sample of senders.

In order to motivate our sender, we offered all psychology students at our University who were taking a social psychology module a substantial academic reward if they participated as witnesses in a lie detection study and were the most convincing of all senders. Four undergraduate females volunteered. Each of them was shown two film sequences depicting criminal actions (S1 and S2). After watching each of these sequences, senders were instructed to work out a deceptive version (D) and a truthful one (T) of the sequence. They were left ten minutes to create each version, and were video recorded as they made their statements –a free narrative account no less than two minutes long. Thus, each sender produced four statements: a deceptive account of the first sequence (S1D), a truthful account of that same sequence (S1T), a deceptive account of the second sequence (S2D) and a truthful account of that second sequence (S2T). A pilot study was conducted with a few undergraduates in order to choose the most convincing liar for the main study[4]. All four candidates received the advertised reward for their participation.

The four performances of the sender who was chosen were edited and shown to 121

[3]In the Spanish National Police Force there are officers (two degrees: *policías* and *oficiales de policía*), subinspectors, inspectors, chief inspectors (*inspectores jefe*), superintendents (*comisarios*), and chief superintendents (*comisarios jefe*). Normally, a superintendent is in charge of a police station, and an inspector is in charge of a group of police officers. To become an officer it is necessary to study the Basic Level (*Escala Básica*) at the Police Academy. Applicants are required to have completed their primary education, as well as to pass a competitive examination. To become an inspector it is necessary to study the Executive Level (*Escala Ejecutiva*) of the Police Academy. To enter the Executive Level applicants (lay people, that is, non-officers) must be 28 or younger, must have studied at the University, and must pass a competitive examination. There is, however, another way to access the Executive Level: Police officers with a given number of years of on-the-job experience may apply for promotion to police inspectors; if their application is successful, they are then sent to the Police Academy and enter the Executive Level. These students are normally older than the former, and have long experience as officers. In either case, completing the Executive Level takes two years at the Academy plus a practical-training year at a police station. Our "police officers" were novice (i.e., young and inexperienced) students of the Executive Level in their second year at the Academy. Larger differences would be expected using very experienced officers (which were unavailable at the time data were collected), but Garrido et al. (1997, 2002) found that the tendency to judge statements as deceptive was stronger among the same novice officers used in this study than among the undergraduate students. The military-like kind of life officers have at the Police Academy may be responsible for the effectiveness of such a brief socialization process, thus accounting for the differences between them and the students which were found by Garrido and his colleagues.

[4]The main study, reported by Garrido, Masip, Herrero and Tabernero (1997) (conference presentation) and Garrido, Masip, and Herrero (2003) (paper under review), compared police officers' and lay people's ability to detect truthful and deceptive statements. In order to ensure that credibility judgments were not obvious, so that differences between more skilled and less skilled groups could emerge, a liar was chosen who, according to the ratings by the participants in the pilot study, was relatively good at deceiving.

police officers and 146 psychology students. Each participant watched two statements: one based on S1 and the other based on S2. These statements could be both truthful (31 police officers and 38 undergraduates were allocated to this condition), both deceptive (29 officers and 40 undergraduates), truthful the first to be shown and deceptive the second to be shown (31 officers and 36 students) or deceptive the first and truthful the second (30 officers and 32 students). All police officers allocated to the same experimental condition were in the same class in the police academy at the moment the experimental session was carried out; allocation of officers to their classes is based on an alphabetical criterion. Allocation of undergraduate students to the experimental conditions was made randomly. Since the number of officers per classroom was not the same across all classrooms, and some students failed to attend their sessions and/or came to a session different from the one they had been assigned to, there were some small variations in size across the experimental groups.

After watching each of the two performances of the sender, observers were given a few minutes to complete a questionnaire. One of the items asked them whether they thought the sender had lied or told the truth. Another item asked observers whether they had reached their conclusion early, as they started to see the sender's performance (Moment 1), at the middle part of that performance (Moment 2), or at the final moment (Moment 3).

Results

Hypothesis Testing

Data were analyzed separately for S1 and S2. Two stepwise backward hierarchical loglinear analyses were performed using SPSS 9.0. The variables introduced were value of truth of the statement (truthful / deceptive), observers' occupation (police officer / undergraduate), the hit / miss variable, and the moment observers made their decision (Moment 1 or 2 / Moment 3)[5]. Both a significant association among occupation and moment –as predicted in hypothesis 1– and a value of truth X moment X hit / miss –in the way predicted in Hypothesis 2– would be expected to emerge for both statements. In addition, concerning the first three variables, we expected to find results similar to those reported by Garrido et al. (2003; see also Masip, 2002), which were based on these data.

Concerning S1, k-way effect tests showed the fourth-order interaction was of no relevance, likelihood-ratio chi-square: $\chi^2_{(1)} = 0.04$, $p = .846$, but there were substantial third-, second-, and first-order effects, respectively: $\chi^2_{(4)} = 18.05$, $p = .001$; $\chi^2_{(6)} = 69.54$, $p = .000$; $\chi^2_{(4)} = 9.68$, $p = .046$. Something similar was found for S2, respectively: $\chi^2_{(1)} = 1.59$, $p = .220$; $\chi^2_{(4)} = 22.80$, $p = .000$; $\chi^2_{(6)} = 22.76$, $p = .001$; and $\chi^2_{(4)} = 28.83$, $p = .000$. The best model for S1 comprised three interactions: Occupation X Value of Truth X Hit/Miss (police officers made more errors when judging truthful statements than when judging the deceptive; this effect was presented and discussed by Garrido et al., 2003), Value of Truth X Moment X Hit/Miss, and Occupation X Moment (these interactions are discussed briefly). This model had an excellent goodness of fit: its likelihood-ratio chi-square was $\chi^2_{(3)} = 0.24$, $p = .971$, and the greatest standardized residual had an absolute value of 0.29. The best model for S2 was somewhat simpler, comprising only the two third-order interactions also included in the S1 model, that is, Occupation X Value of Truth

[5] Few judgmental decisions were made at Moment 1: only 77 (28 of which came from the first statement [S1] and 49 from the second [S2]). At moments 2 and 3 the number of judgments was virtually the same (N = 231 at Moment 2, N = 223 at Moment 3); this was so both in the first statement (122 at Moment 2, 117 at Moment 3) and in the second statement (109 at each moment), and these judgments were much more numerous than those of the first moment. This small frequency of initial judgments turned out to be a limitation: by looking at the expected frequencies in the contingency tables where all the variables to be introduced in the loglinear analysis were crossed, it became evident that, in moment one, these were too small to conduct the analysis. Therefore, in order to calculate the statistics moments 1 and 2 were grouped and taken together. Thus, this variable had two categories: Moments 1 and 2 v. Moment 3.

X Hit/Miss and Value of Truth X Moment X Hit/Miss. This model had a likelihood-ratio chi-square of $\chi^2_{(4)} = 4.93$, $p = .294$, and the greatest standardized residual had an absolute value of 0.85.

In order to examine the specific contribution of each effect to the fit of the model, attention was paid to partial association tests and parameter estimates. In Table 1 this information is summarized for all those effects which either approached significance or were significant, whether in S1 or in S2. The direction of effects can be observed in Appendix 1, where the presentation model suggested by Tabachnick and Fidell (1996) was used.

Concerning the occupation, value of truth, and hit / miss variables, Garrido et al.'s (2003) results were here replicated with some minor nuances. In particular, the most relevant effect (the three-way interaction) was found again and, furthermore, it did not interact with the moment observers decided their judgment. Thus, the introduction of that variable in the analyses did not substantially alter the former results. Since they were already discussed elsewhere (Garrido et al., 2003) and are not the main focus of the present report, they will not be discussed here again.

Instead, our focus in the present paper centers on those effects involving the moment variable. A certain tendency was found in S1 to make the decisions at moments 1 and 2 (56.18 % of judgments) instead of making them at Moment 3 (43.82 %). Something similar, although the trend was clearer, happened in S2 (percentages were, respectively, 59.19 % and 40.82 %). This effect could be due to having added the number of decisions made at Moment 1 to those made at Moment 2, since in both statements the latter had a frequency that was quite similar to that of Moment 3 decisions.

The *first hypothesis* predicted that police officers would hurry and make their judgment earlier than the undergraduates. Therefore, an association between being an officer and deciding at moments 1 and 2, and between being a student and deciding at Moment 3 would be expected. However, the Occupation X Moment association was not significant in S2. In S1 it did not reach statistical significance either, as indicated by the two measures used to explore the individual effects (although it was close to significance: $\chi^2_{(1)} = 3.24$, $p = .072$; $z = -1.92$), but the program retained the effect while searching for the best model during the stepwise procedure (the associated change

Table 1. *Partial Association Tests And Parameter Estimates (In Absolute Values) Of The Effects That Had A Relevant Contribution To The Fit Of The Model, Either In S1 Or In S2.*

Effects	Partial association				Parameter estimates*			
	S1		S2		S1		S2	
	$\chi^2_{(1)}$	p	$\chi^2_{(1)}$	p	$\lvert\lambda\rvert$	$\lvert z\rvert$	$\lvert\lambda\rvert$	$\lvert z\rvert$
Third-order effects								
Occupation X Value of Truth X Hit/Miss	8.13	.004	7.04	.008	.20	2.66	.17	2.44
Value of Truth X Moment X Hit/Miss	11.40	.001	12.65	.000	.24	3.12	.24	3.58
Second-order effects								
Occupation X Moment	3.24	.072	0.45	.832	.15	1.92	.02	0.25
Value of Truth X Moment	1.39	.239	12.18	.001	.03	0.42	.19	2.82
Value of Truth X Hit/Miss	49.63	.000	3.87	.049	.47	6.20	.10	1.47
Moment X Hit/Miss	12.25	.001	0.99	.321	.21	2.80	.09	1.32
First-order effects								
Moment	4.09	.043	9.88	.002	.04	.55	.13	1.96
Hit/Miss	3.16	.076	16.11	.000	.16	2.18	.19	2.79

* In absolute values. To examine the direction of effects see Appendix 1.

likelihood-ratio chi-square was $\chi^2_{(1)} = 5.48$, $p = .019$). In any case, the effect was the opposite to what was expected: police officers did not tend to make their decisions earlier than the undergraduate students, but later (see Appendix 1): While in S1 48.76 % of police officers decided at moments 1 and 2 in comparison with the remaining 51.24 %, who decided at Moment 3, 62.33 % of undergraduates decided during the early moments and only 37.67 % of them did so at Moment 3. In S2 the associations failed to reach significance, but they pointed in the same direction. In summary, our first prediction did not receive empirical support. If there was any occupational group which acted hastily in making their credibility judgments it was not the officers, but rather the undergraduate students.

The *second hypothesis* predicted an interaction between the value of truth of the statement, the decision-making moment, and the correctness of the credibility judgment (hit or miss), in the sense that deceptive statements would be more accurately detected earlier in the statement than later on, while truthful statements would be judged with higher accuracy at the final moment rather than at the beginning of the statement. To begin with, it should be mentioned that some second-order effects were substantial. In S2, the interaction Moment X Value of Truth indicates that, when judging the false account (S2D), the decision was made basically at the beginning (69.85 % of cases); this did not happen when judging S2T (48.09 %). In S1 it was the Moment X Hit/Miss interaction that was relevant: the decisions made at the beginning of the statement were accurate more often than those made at the final moment. But both of these effects were qualified by the higher-order value of Truth X Moment X Hit/Miss interaction, which lent support to our second hypothesis. When judging the deceptive statements an association was found both in S1 and S2 between making the decision early (moments 1 and 2) and guessing right, as well as between deciding at the final moment and judging wrongly. An opposite tendency became apparent when judging the truthful statements (see Appendix 1). As stated before, this effect was one of the components of the final model in both analyses: the one concerning S1 and the one concerning S2.

Although this interaction was significant it would be interesting to analyze whether, in an absolute sense, there was a significant decrease in accuracy when judging deceptive statements at Moment 3 in comparison with the early moments, as well as whether the increase in accuracy for the truthful statements was significant too. In order to examine these effects, individual chi-square analyses were performed to examine the associations among the hit/miss and the moment variables separately for truthful and deceptive accounts. The results of these analyses are summarized in Table 2. It is apparent that the predicted decrease for the deceptive statements was significant. However, the increase in accuracy for truthful statements across time was not found, although a marginally significant trend in the predicted direction was apparent in S2T.

In conclusion, our second hypothesis was supported by the data. There was a relative decrease in accuracy over time when judging deceptive statements, and a relative increase when judging the truthful ones. However, in an absolute sense, although the decrease when judging deceptive statements was significant, the increase of judgmental accuracy for truthful statements did not reach statistical significance.

Early Lie Bias

It is worth noticing that, aside from the idiosyncrasies of each statement, the data reported here show that, in general, a strong initial bias toward making judgments of deceptiveness was apparent. There were large differences between observed frequencies of hits and misses at moments 1 and 2, both when the statements were deceptive (many more hits than misses) and when they were truthful (more misses than hits) (see Table 2). At Moment 3 these differences were severely reduced or reversed.

Table 2. *Moment X Value Of Truth X Hit/Miss Contingency Tables, And Chi-Square Analyses For Truthful And Deceptive Statements.*

Statements	Moment		χ^2 (1)	*p*
	Moments 1 and 2	Moment 3		
Sequence 1 (S1)				
Deceptive (S1D)				
Hit	62 (1.8)	24 (-2.1)	22.53	.000
Miss	13 (-2.5)	32 (2.9)		
Truthful (S1T)				
Hit	19 (0.2)	14 (-0.2)	0.10	.747
Miss	56 (-0.1)	47 (0.1)		
Sequence 2 (S2)				
Deceptive (S2D)				
Hit	74 (1.1)	19 (-1.6)	12.13	.000
Miss	21 (-1.6)	21 (2.4)		
Truthful (S2T)				
Hit	30 (-0.8)	42 (0.8)	2.98	.084
Miss	33 (0.9)	25 (-0.9)		

Although the increased number of initial judgments of deceptiveness seemed to be more evident when the statements were deceptive, the differences between the number of truth and lie judgments were statistically significant not only in that case, $\chi^2_{(1)} = 58.82$, $p = .000$, but also when statements were truthful, $\chi^2_{(1)} = 11.59$, $p = .001$. On the contrary, at Moment 3 there were no significant differences between judgments of truthfulness and judgments of deceptiveness, either when statements were deceptive, $\chi^2_{(1)} = 0.67$, $p = .414$, or when they were truthful, $\chi^2_{(1)} = 1.53$, $p = .212$. All these results indicate that lie judgments were more numerous when deciding at the beginning of statements than when the decision was made at the end. Indeed, a Moment (moments 1 and 2 / Moment 3) X Judgment (judgment of truthfulness/ of deceptiveness) *Chi-square* analysis was significant, $\chi^2_{(1)} = 25.66$, $p = .000$.

These tendencies cannot be accounted for by the differences between police officers and undergraduates in terms of the kind of judgment each group tended to make. As reported elsewhere (Garrido et al., 1997, 2003), police officers' tendency to deem statements as deceptive was stronger than non-officers', $\chi^2_{(1)} = 9.57$, $p = .002$. This tendency was significant not only at moments 1 and 2, $\chi^2_{(1)} = 8.81$, $p = .003$, but also at Moment 3, $\chi^2_{(1)} = 4.05$, $p = .044$, and the difference between truth and lie judgments at moments 1 and 2 was significant not only among officers, $\chi^2_{(1)} = 51.72$, $p = .000$, but also among the students, $\chi^2_{(1)} = 18.90$, $p = .000$, while at Moment 3 the difference in frequency of truth and lie judgments was only marginally significant among police officers, $\chi^2_{(1)} = 2.95$, $p = .086$, and was completely non-significant among the undergraduates, $\chi^2_{(1)} = 1.26$, $p = .261$. Therefore, the differences between the officers and the students cannot account for that trend towards judging statements as deceptive fundamentally at the beginning of the sender's performance. In addition, in this regard it is important to keep in mind that the officers, whose tendency to judge the statements as deceptive was stronger than undergraduates', displayed a certain propensity to make their judgments later than the students. Also, a backward stepwise hierarchical loglinear analysis was performed to examine the relation between Occupation, Moment, and Judgment (truth / lie judgment). If the association between moment and judgment were moderated by the observers' occupation, then k-way effect tests would yield significant results for k = 3, and the analysis would not continue beyond the saturated model (Occupation X Moment X Judgment). However, the null hypothesis that third-order effects were zero was supported, likelihood-ratio chi-square: $\chi^2_{(1)} = 0.48$, $p =$

.488, and the best model comprised the Occupation X Judgment interaction, partial $\chi^2_{(1)} = 12.65$, $p = .000$, $|\lambda| = .17$, $|z| = 3.47$ (which reflects the police's tendency to judge the statements as deceptive), the Occupation X Moment interaction, partial $\chi^2_{(1)} = 5.49$, $p = .019$, $|\lambda| = .12$, $|z| = 2.41$ (which reflects the aforementioned tendency among police officers to make their judgment later than the undergraduates, which in this analysis was clearly significant), and the judgment X moment interaction, partial $\chi^2_{(1)} = 28.60$, $p = .000$, $|\lambda| = .25$, $|z| = 5.27$ (which reflects the tendency we are discussing to make lie judgments at the beginning of the statements but not at the end of them). This model had an adequate goodness of fit, likelihood-ratio chi-square: $\chi^2_{(1)} = 0.48$, $p = .488$; the greater standardized residual had an absolute value of 0.36.

In summary, regardless of whether statements were truthful or deceptive, early judgments were primarily lie judgments. Later on, at Moment 3, truth and lie judgments were more balanced. Police officers' tendency to make lie judgments cannot account for this effect.

Discussion

Hypothesis 1: The moment officers and non-officers made their decision

Hypothesis one, which predicted that police officers would make their decision early in the statement and that students would decide later, was not supported. Differences were contrary to what was expected (i.e., there was a tendency among officers to decide at the final moment and among the students to do so at the beginning), and the effect was retained in some of the loglinear analyses. Maybe officers, due to their awareness that information gathered from witnesses is important, were more attentive and did not decide until they had collected the information, while students were less thorough and hastened their decision. Nevertheless, officers' bias to judge truthful statements as deceptive (Garrido et al., 1997, 2003) suggests that they were either incapable of gathering the relevant information in order to make their veracity judgments, or they did not make correct use of the information they collected.

Hypothesis 2: The influence of moment on accuracy

Hypothesis 2 was partially supported. Deceptive statements tended to be judged accurately at the beginning and inaccurately at the end. Probably, the greater amount of misleading information at Moment 3 in comparison to moments 1 and 2 made observers less accurate at detecting deception, since that information serves to (a) give an impression of being honest, and (b) make the lie plausible and, hence, credible (Buller & Burgoon, 1996, 1998; Leekam, 1992). This leads to truth judgments. Clues to deceit will increase later in time as well, but research shows that untrained observers are not good at using such clues; instead, they often base their judgments on invalid indicators (DePaulo et al., 1985; DePaulo, Zuckerman, & Rosenthal, 1980b; P. DePaulo et al., 1989; Ekman, 1989; Vrij, 1998, 2000; Vrij & Winkel, 1993).

However, for truthful accounts the frequency of observed hits and misses at moments 1 and 2 in comparison with Moment 3 did not depart significantly from what was expected. In other words: observers' judgments were more or less equally accurate at any moment in time. This is at odds with our second hypothesis, which predicted an increase in accuracy across time. In any case, as stated above, hypothesis two was just an exploratory hypothesis, and the present findings are only preliminary. A study with a large number of senders is currently being conducted to replicate these findings.

Early Lie Bias

Our results show a strong initial lie bias for our observers. This was so among both students and officers, despite the finding reported by Garrido et al. (1997, 2003) that, overall, officers were more prone to make judgments of deceptiveness than non-officers. Thus, while an initial accuracy level common for both truthful and deceptive statements and close to chance probability could be expected, since at the beginning of the sender's performance the amount of both accurate and misleading information was similar for all statements and similarly scarce, a strong tendency to judge statements as false was found at that point in time. Thus, when observers decided early in time they tended to

make judgments of deceptiveness. A possible explanation for this initial bias may lie on our sender's physical appearance. Research indicates that people's facial appearance may influence social perceivers' impressions of sincerity (e.g., Berry & Brownlow, 1989; Berry & McArthur, 1985; Zebrowitz & Montepare, 1992; Zebrowitz, Voinescu, & Collins, 1996). The only available information at the beginning of the statements was the sender's appearance. Thus, it is possible that initial credibility judgments were influenced by by our sole sender's appearance. If our sole sender had a facial appearance that fitted the stereotype of a liars' face, the intriguing initial lie bias could be due to that factor. This possibility was investigated in Experiment 2.

Disproportion Of The Number Of Decisions Made At Different Moments

The small number of decisions made at Moment 1 in comparison with those made at moments 2 and 3 may be due to the subjective nature of the distinction between Moment 1 and Moment 2. There were no "markers" of the boundaries between the different moments. Observers who decided at Moment 3 were probably those who waited until the video presentation was over to judge whether the sender lied or told the truth. However, boundaries between Moment 1 and 2 were not so clear. What for some observers was Moment 1, may have been Moment 2 for others. In addition, the small number of Moment 1 decisions may indicate that, unless observers decided at the very beginning of the video clip (and not at any point within the first third), they generally said they decided at Moment 2. Replications using clear "markers" to separate the time periods of interest, such as questions by an interviewer (answer to the first question: Moment 1, answer to the second question: Moment 2, and so on), an acoustic signal (e.g., before the first beep: Moment 1, first to second beep: Moment 2, etc.), or a time display on the TV screen (e.g., first minute: Moment 1, second minute: Moment 2, and so on) would help us clarify that issue. Work in progress is addressing this point.

Experiment 2

As noted above, in the experiment just described a strong lie bias was found for early judgments. An interesting question is why observers showed that bias. Actually, the information they had at that early point in time was scarce, apart from the sender's physical appearance. Is it possible that a person's appearance influences credibility judgments? There is evidence indicating that this could be the case. For instance, one's facial appearance has been found to influence a series of attitudes, behaviors, attributions, and judgments made by others (see reviews by Alley, 1988; Berry & Zebrowitz, 1986; Bruce & Young, 1998; Bull, 1982; Bull & McAlpine, 1998; Bull & Rumsey, 1988; Shepherd, 1989; Zebrowitz, 1997). Thus, might it be possible that there is a social stereotype of the appearance of a liar's face, so that initial credibility judgments are influenced by the extent to which the sender looks like an honest individual or a deceptive one. If so, someone with a liar-looking facial appearance would be judged as deceptive early in his or her statement, but perhaps the availability of information provided by the sender as time goes by can reduce that initial tendency.

Zuckerman, DeFrank, Hall, Larrace, and Rosenthal (1979), found what they termed a demeanor bias in their senders: some were consistently judged as honest and some as deceptive, regardless of whether they lied or told the truth. The existence of a demeanor bias has been confirmed by later research conducted by Bond, Kahler, and Paolicelli (1985). As conceptualized by Zuckerman et al. (1979), that bias would depend on some internal characteristics influencing the sender's perceptible demeanor, which, in turn, would determine observers' ratings. Indeed, some authors have tried to see the influence of some personality traits and social skills of the sender upon observers' credibility judgments (e.g., Geis & Moon, 1981; Miller, deTurck, & Kalbfleisch, 1983; Riggio & Friedman, 1983; Riggio, Tucker, & Widaman, 1987; Riggio, Tucker, & Throckmorton, 1987; Vrij, 1992; Vrij & Winkel, 1993), assuming that these traits and skills influence in some way the behavior displayed by the communicator (for empirical tests of this assumption see Riggio, Tucker, & Widaman, 1987; Vrij, Akehurst, & Morris, 1997). However, as Bond and Robinson (1988) suggest, it may be the case that "these biases originate in fixed features of the mien, an innocent- or guilty-looking visage" (p. 304). If this were the case, then the biased judgments of credibility would depend

directly upon the sender's appearance, instead of depending on some personality traits or social skills that influence behavior. Remember that in Exeriment 1 we used only one sender. If that sender had a face that fits the social stereotype for the face of a liar, then her appearance could have been responsible for observers' initial lie-bias. Later in time, however, behavioral information drawn from the sender's behavior may have reduced that bias. Specifically, the misleading information provided in the false stories reduced judgments of deceptiveness. Unlike us, Zuckerman et al. (1979) used series of 15-second videotaped segments, too brief a time period to find a reduction in the demeanor bias. That is, all judgments in Zuckerman et al.'s study were made at what in our experiment was Moment 1 (or, at best, what our observers regarded as Moment 2); that is probably why they found such a strong demeanor bias. One of the aims of the present, follow-up experiment was to check whether an initial lie bias is found again if the sender's face is different from that of study one.

Here we took two of the statements used in Experiment 1 (S1T and S1D). Those statements were presented via audio, while a still face of a young woman, supposedly the one making the statement, appeared on a TV screen. That face could be of the same sender as in Experiment 1, or of two other senders. Comparing the initial accuracy for truthful and deceptive statements for the several purported senders enabled us to check whether the initial lie bias found in Experiment 1 was due to the facial appearance of the sender used in that study.

Method

Participants

The sender was one female undergraduate student of psychology at a Spanish University. Still images of two other senders, also females of similar ages, where used. Observers were 224 police officers studying to become police inspectors at the Police Academy.

Procedure

The procedure used to create the statements is described in the method section of Experiment 1. Here in Experiment 2, audio recordings of S1T and S1D were presented. The decision to select the two versions of only one original sequence was prompted by the need to use a limited number of participants[6]. Sequence 1 was chosen because S1T and S1D were entirely different from one another, while S2D was a variation of S2T where central details were changed to make it deceptive. Since, as we shall see later, each participant would have to judge both statements, these had to be entirely different from one another. If we had used S2, judgments for the second statement would not have been independent from those for the first.

Video clips were edited where the audio recordings of S1T and S1D were coupled with a still image of the witness who purportedly

[6] Notice that in Experiment 1 four groups of observers were used (each of these was in turn comprised of a subgroup of undergraduates and another one of police officers). In Experiment 2 three different still faces had to be shown while the same words as in Experiment 1 were heard. This makes 12 groups, too large a number of samples. Therefore, only the statements based on one of the original video sequences were taken for this experiment. This was not a problem, since the strong tendency found in Experiment 1 to make judgments of deceit at moments 1 and 2 but not at Moment 3 was evident for both S1, χ^2 (1) = 10.28, p = .001, and S2, χ^2 (1) = 17.51, p = .000. Also, police officers were taken as observers in Experiment 2 not only because they were available at the time data were to be collected for that experiment, but also because it increases the external validity of the findings when it comes to generalizing them to real criminal cases. In addition, using only officers as observers was not problematic since, as reported above, in Experiment 1 the tendency to judge statements as deceptive at the beginning of the sender's performance was statistically significant among officers, while at moment three that trend had at best a marginal significance. In fact, chi-square analyses made on the data of Experiment 1 to examine the relation between moment (1 and 2 v. 3) and judgment (truthfulness / deceptiveness judgment) were significant for both undergraduate students, χ^2 (1) = 13.24, p = .000, and police officers, χ^2 (1) = 15.73, p = .000, and this was so not only for S2, students: χ^2 (1) = 5.42, p = .020, officers: χ^2 (1) = 14.11, p = .000, but also for S1 which, as mentioned above, was the statement chosen to be used in Experiment 2, students: χ^2 (1) = 7.78, p = .005, officers: χ^2 (1) = 5.09, p = .024.

had made the statement. This image was of the same sender shown in Experiment 1, or one of two other senders who initially volunteered to participate in our study and had made their statements. These pictures were taken from the tapes of their statements. All senders, as shown in the still images employed, faced the camera and displayed a neutral facial expression.

Groups and number of observers per group are shown in Table 3. Again, all those police officers who attended their lectures in a given classroom were allocated to the same experimental group. As mentioned in Experiment 1, allocation of officers to their classrooms is based on an alphabetical criterion. The judgmental sessions were similar to those described in Experiment 1. Statements were presented to observers via a videotape connected with a TV monitor. Observers completed the same questionnaires as in Experiment 1, although some additional questions were added. One asked observers how attractive they found the sender. Answers were collected on a continuous scale from 1 (very unattractive) to 7 (very attractive). Another question asked observers how old they thought the sender was. These two questions were at the end of the questionnaire.

Results

Manipulation Checks

If we are to analyze how facial appearance influences credibility judgments we must first make sure that our facial stimuli are different from each other in some characteristics likely to influence social judgments. Two such characteristics are age (e.g., Montepare & Zebrowitz, 1998) and attractiveness (e.g., Alley & Hildebrandt, 1988; Zebrowitz, 1997).

Age

Observer's ratings of targets' age were 23.47 years for Face A, 24.36 for Face B, and 22.76 for Face C (which was the face of the sender we used in Experiment 1), $F_{(2,221)} = 6.75$, $p = .001$. Post-hoc Fisher's LSD tests showed that Face B was judged as significantly older than Face A, $p = .034$, and Face C, $p = .000$, but Face A was not perceived as significantly older than Face C, $p = .111$.

Attractiveness

Observers also rated the degree of attractiveness of the faces. Ratings were 4.76 for Face A, 3.98 for Face B, and 4.59 for Face C, $F_{(2,221)} = 21.79$, $p = .000$. Fisher's LSD tests showed that Faces A and C were perceived as similarly attractive, $p = .209$, but Face B was judged as less attractive than faces A and C, both ps = .000.

Credibility Judgments

In Experiment 1 a lie bias, which was greater among police officers than among undergraduates (see Garrido et al., 1997, 2003), was found. That bias decreased as observers made their decision about the witness's veracity later in time. Unlike Experiment 1, in this study visible dynamic cues displayed by the sender (e.g., her gestures and body movements) were absent, her facial appearance was manipulated, and all the observers were members of the Spanish National Police Force. In Experiment 2 we addressed the following questions: First, whether in these circumstances a lie bias also appears; second, whether this bias decreases across time; and third, whether this is dependent upon the sender's facial appearance, that is to say: (a) whether the lie bias is apparent for any of the faces (i.e., that of Experiment 1) but not for the others, and

Table 3. *Groups And Number Of Observers Per Group In Experiment 2.*

	Faces		
Pairs of statements	A	B	C (same as in Exp. 1)
S1T – S1D	38	42	36
S1D – S1T	40	42	26

(b) whether the decrease of that bias over time happens when any of the still faces is presented but not when the others are presented.

In order to find an answer to these questions we conducted two backward stepwise hierarchical loglinear analyses, one for the truthful statement and another one for the deceptive one (participants who had judged each statement were exactly the same; the order of the truthful and the deceptive statement was counter-balanced, as shown in Table 3). The variables which were introduced in the analyses were observers' veracity judgment (truthfulness / deceptiveness judgment), the face (A / B / C), and the moment observers said they made their decision concerning the witness's credibility (Moments 1 and 2 / Moment 3)[7].

For the *truthful statement*, k-way effect tests supported the null hypothesis that third-order effects were equal to zero, likelihood-ratio chi-square: $\chi^2_{(2)} = 0.77$, $p = .681$, and rejected the hypotheses that second-, $\chi^2_{(5)} = 24.32$, $p = .000$, and first-order effects, $\chi^2_{(4)} = 44.70$, $p = .000$, were zero. The best model comprised two second-order interactions: Judgment X Moment, partial $\chi^2_{(1)} = 15.45$, $p = .000$, and Moment X Face, partial $\chi^2_{(2)} = 8.12$, $p = .017$. In addition, the main effect of judgment had a significant partial chi-square, $\chi^2_{(1)} = 40.38$, $p = .000$. The model had an adequate goodness of fit: likelihood-ratio $\chi^2_{(4)} = 4.11$, $p = .392$; the standardized residual that had a larger absolute value was 0.89. Parameter estimates and their corresponding *z* values are shown in Appendix 2. In addition, Table 4 shows the observed frequencies and the standardized residuals with reference to the independence model corresponding to the two interactions of the final hierarchical model. The judgment effect indicates that, just as in Experiment 1, the frequency of lie judgments (70.72 %) was larger than the frequency of judgments of truthfulness (29.28 %). This was so regardless of the face that was presented, because the Face X Judgment effect was not significant. However, credibility judgments were actually affected by the moment the decision was made, as indicated by the Judgment X Moment interaction: there was an association between making the decision at Moment 1 or 2 and judging the statement as deceptive, and making it at Moment 3 and judging the statement as truthful (see Table 4 and Appendix 2). Thus, a decrease over time of the lie bias was found here as well. This happened regardless of the face that was presented. However, the face had an effect, not upon whether statements were judged as truthful or deceptive, but on the moment the decision was made: those who watched face A tended strongly to make their decision at the beginning of the statement, while those who watched face B tended moderately to decide at the end. The tendency for face C was not significant[8] (see Table 4 and Appendix 2).

For the *deceptive statement* the results were quite similar. K-way effect tests failed to yield significant results with regard to the third-order effect, $\chi^2_{(2)} = 2.76$, $p = .252$, but not with regard to second-, $\chi^2_{(5)} = 35.65$, $p = .000$, and first-order effects, $\chi^2_{(4)} = 57.35$, $p = .000$. The final model comprised exactly the same interactions as for the truthful statement: Judgment x Moment, partial $\chi^2_{(1)} = 11.55$, $p = .001$, and Moment X Face, partial $\chi^2_{(2)} = 18.70$, $p = .000$. The first-order effect of judgment was significant also in this case, $\chi^2_{(1)} = 53.52$, $p = .000$, indicating that judgments of deceptiveness (73.99 %) exceeded judgments of truthfulness (26.01 %). The fit of

[7] Once again, when differentiating between moments 1, 2, and 3, expected frequencies were too small at Moment 1, particularly when judgments of truthfulness were made. Therefore, for both truthful and deceptive statements, moment-1 and moment-2 judgments were taken together and compared with moment-3 judgments.

[8] Two variables in a contingency table (such as Table 4) are related in a cell if the standardized residual in that cell has an absolute value equal to or higher than 1 (Martín, Cabero, & Ardanuy, 1997). Also, two variables are related in a cell such as those of Appendix 2 if the associated *z* value has an absolute value equal to or higher than 1.96 (e.g., Tabachnick & Fidell, 1996).

the model was good, with a likelihood-ratio chi-square of $\chi^2_{(4)} = 3.16$, $p = .531$, and the greater standardized residual had an absolute value of 0.78. As shown in Table 4 and Appendix 2, regardless of the face which was presented an association between making the decision at moments 1 or 2 and judging the statement as deceptive was found, as well as an association between deciding at Moment 3 and judging the statement as truthful. Also, just as in the former case, those who watched face A tended to make their judgment at moments 1 and 2, those who watched face B tended to make it at Moment 3, and the tendency for face C was not significant.

In summary, as was the case in Experiment 1, Experiment 2, for both the truthful (S1T) and the deceptive (S1D) statements found: (a) a lie bias, (b) this bias decreased over time, and (c) neither of these effects was influenced by the purported witness' facial appearance. An influence of the facial appearance upon the moment decisions were made was found as well: both when the statement was truthful and when it was deceptive, face A judgments were made at the beginning of the statement, and face B judgments were made at the end.

Detection Accuracy

In the preceding paragraph the conclusion was drawn that there was a strong tendency to say the sender was lying. This should have an influence on accuracy, so that judgments of the truthful statement should be wrong more often than judgments of the deceptive statement, and the latter should be accurate more often than the former. In addition, we have seen that this lie bias tended to decrease over time. Therefore, the trends towards judging incorrectly of the truthful statement and guessing correctly those of the deceptive statement should decrease over time as well.

To examine those questions two backward stepwise hierarchical loglinear analyses were calculated, one for the first statement that was presented (first presentation) and the other for the second (second presentation) (some observers watched S1T first, and then S1D; other watched the clips in the reverse order; see Table 3). The variables which were

Table 4. *Observed Frequencies And Standardized Residuals With Reference To The Independence Model Corresponding To The Judgment X Moment And The Moment X Face Interactions For The Truthful Statement And The Deceptive Statement.*

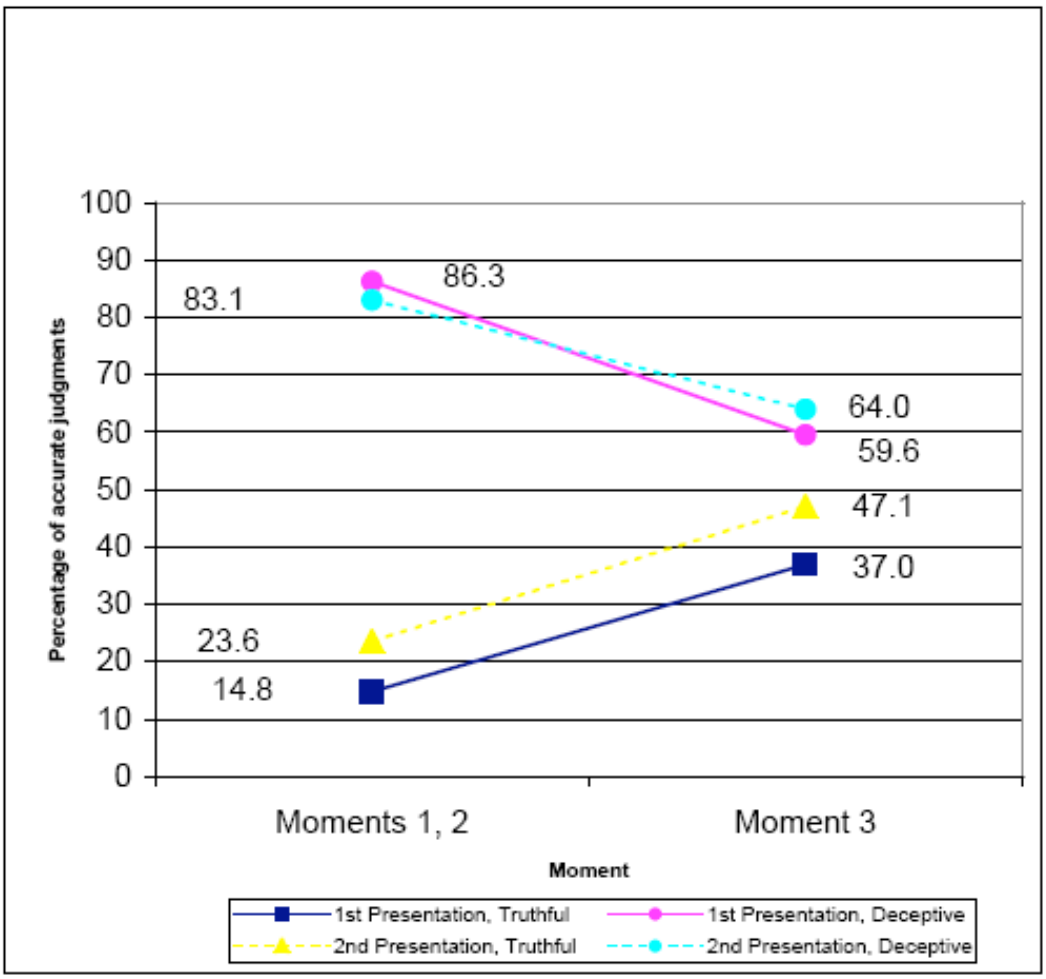
JUDGMENT X MOMENT

Judgment	Moment			
	Moments 1 and 2		Moment 3	
Truthful Statement				
Judgments of truthfulness	21	(-2.2)	43	(2.3)
Judgments of deceptiveness	95	(1.4)	62	(-1.5)
Deceptive Statement				
Judgments of truthfulness	18	(-2.2)	40	(2.3)
Judgments of deceptiveness	98	(1.3)	67	(-1.4)

MOMENT X FACE

Moment	Face					
	A		B		C	
Truthful Statement						
Moments 1 and 2	50	(1.4)	38	(-0.9)	29	(-0.5)
Moment 3	28	(-1.5)	46	(1.0)	32	(0.6)
Deceptive Statement						
Moments 1 and 2	56	(2.4)	30	(-2.1)	31	(-0.2)
Moment 3	22	(-2.5)	54	(2.2)	31	(0.3)

Figure 1. Accuracy rates on judging the credibility of truthful and deceptive statements at moments 1 and 2, as well as at Moment 3, for the first and the second presentations of Experiment 2.



introduced in the analyses were Moment (1 and 2 *v.* 3), Value of Truth, and Hit / Miss. In both cases the k-way effect test indicated that the third order interaction was significant: likelihood-ratio chi-squares were: $\chi^2_{(1)}$ = 17.54, p = .000, for the first presentation, and $\chi^2_{(1)}$ = 11.83, p = .001, for the second. Consistent with these results, in neither case did the process continue beyond the saturated model. As shown in Appendix 3, the Value of Truth X Hit/Miss interaction was substantial. Both for the first presentation, partial $\chi^2_{(1)}$ = 51.29, p = .000, and for the second presentation, partial $\chi^2_{(1)}$ = 36.57, p = .000, judgments of the truthful statement tended to be wrong, and those of the deceptive statement tended to be accurate. In fact, the percentage of accurate judgments of the truthful statement was only 25.22 % in the first presentation, and 34.91 % in the second; the corresponding values for the deceptive statement were 72.22 % and 74.78 % (see Figure 1). Therefore, the lie bias had a strong effect upon accuracy. However, this effect was influenced by the moment the decision was made, as indicated by the Moment X Value of Truth X Hit/Miss effect in both analyses (see Appendix 2): it was stronger at the beginning of the statement than at the final moment, as shown in Figure 1. In fact, when the decision was made at the initial moments, there was a

larger proportion of accurate judgments for the deceptive account than for the truthful, for the first presentation: $\chi^2_{(1)} = 23.11$, $p = .000$; for the second presentation: $\chi^2_{(1)} = 25.09$, $p = .000$. However, when the decision was made at the end, although the proportion of accurate judgments of the deceptive account was still somewhat larger than that of the truthful, one this difference was not statistically significant, for the first presentation: $\chi^2_{(1)} = 3.63$, $p = .057$; for the second presentation: $\chi^2_{(1)} = 1.14$, $p = .285$. These effects are clearly shown in Figure 1.

Similarly, in the early moments the proportion of errors upon judging the truthful statement was significantly larger than the proportion of accurate judgments for both the first presentation: $\chi^2_{(1)} = 30.31$, $p = .000$, and for the second presentation: $\chi^2_{(1)} = 15.29$, $p = .000$. On the contrary, the proportion of errors at judging the false statement was smaller than the proportion of correct judgments for both the first presentation: $\chi^2_{(1)} = 26.84$, $p = .000$, and for the second presentation: $\chi^2_{(1)} = 28.45$, $p = .000$. In the final moment the differences were in the same direction, but the residuals and, consequently, the significance, decreased with respect to the initial moments: truthful statement in first presentation: $\chi^2_{(1)} = 3.63$, $p = .057$, in second presentation: $\chi^2_{(1)} = 0.18$, $p = .674$; deceptive statement in first presentation: $\chi^2_{(1)} = 2.12$, $p = .145$ and in second presentation: $\chi^2_{(1)} = 3.92$, $p = .048$.

In conclusion: the overall lie bias made accuracy for the deceptive statement higher than accuracy for the truthful one. As this bias decreased over time, so did the tendency to be more accurate when judging the deceptive statement than when judging the truthful one.

Discussion

The second experiment was planned as a follow-up study after the first one. Some experimental conditions of Experiment 1 were changed to check whether Experiment 1 results concerning the moment changed. Thus we hoped to identify the factors determining such results. More specifically, the questions addressed were: (a) Does the lie bias found in the first experiment hold true when dynamic visible cues (gestures and body movements) are suppressed from the videotaped statement?, (b) Does this bias decrease as observers decide later in time, just as happened in the previous experiment?, and (c) Does the existence of the lie bias depend upon the sender's facial appearance, so that a demeanor bias is in operation? To find an answer to those questions a truthful and a deceptive statement of Experiment 1 were presented to observers in their audio format, accompanied by a still image of the person who, supposedly, had enacted the statements. This image could be of the sender who had been used in Experiment 1, or of one of two other young women. Observers had to indicate whether each statement was truthful or deceptive and at what moment they had come to their conclusion on the veracity of the statement, at the beginning (Moment 1), middle (Moment 2) or end (Moment 3) of the videotaped statement.

Analyses were performed to explore the relationships between credibility judgments, the moment at which they were made, and the still face being shown. Results indicate that, overall, the lie bias found in Experiment 1 appeared in Experiment 2 as well: both when judging the truthful statement and when judging the deceptive one, the number of deception judgments was substantially larger than the number of truthfulness judgments. Consequently, there was an association between judging the truthful statement and doing so incorrectly, and between judging the deceptive statement and guessing it correctly. Now, did this effect hold true for the various faces, or only for some of them? And, did it depend in any way on the moment when the judgment was made? Our data indicate that: (a) the decision-making moment had an influence on judgments: the lie bias decreased as time went by, and (b) the witness' facial appearance did not affect credibility judgments.

The Influence Of Moment Upon Veracity Judgments And Accuracy

When the decision was made at moments 1 or 2, a tendency to say statements were deceptive was found; this tendency decreased when the decision was made at Moment 3. This made early accuracy rather high for the deceptive statement and rather

poor for the truthful one, but at Moment 3 these differences had lost significance, although deceptive-statement judgments continued being slightly more accurate than truthful-statement judgments[9]. It is interesting that the Judgment X Moment interaction was significant not only in the loglinear analysis calculated for the false statement, but also in the one calculated for the truthful. Remember that in Experiment 1 the predicted increase in accuracy for truthful statements failed to reach significance, that is, the frequency of truthfulness judgments at Moment 3 was not higher than its frequency at moments 1 and 2. However, in the present experiment, regardless of the still face being shown, accuracy for truthful statements increased significantly over time[10]. Also, the effect already detected in Experiment 1 consisting of a decrease when judging deceptive statements was found here as well.

What reason may account for the fact that the formerly predicted increase in accuracy for the truthful statements did not emerge in Experiment 1 but did appear in Experiment 2? This prediction was based on the assumption that, at the end of the statement, there would be a maximum amount of available accurate information when truthful statements were being presented, while at this same moment the misleading information would reach its maximum when deceptive statements were being presented. This would result in a progressive increase in accuracy over time when judging the truthful statements, coupled with a decrease when judging the deceptive ones.

With this is mind, we should point out that research has shown that verbal cues are the most useful when it comes to making credibility judgments, while nonverbal indicators (gestures and movements) are in general the most misleading (see meta-analyses by DePaulo, Zuckerman, & Rosenthal, 1980a; DePaulo et al., 1985; Kalbfleisch, 1985; Zuckerman et al., 1981). If both kinds of information (i.e., visual and verbal indicators) are presented at the same time, observers probably do not pay much attention to verbal cues, which are the most useful, attending instead to the visual information, which is the most misleading. This would be consistent with the distraction hypothesis, the information overload hypothesis, or the situational familiarity hypothesis. The *distraction hypothesis* posits that visual cues would distract observers from processing verbal and vocal information (Maier & Thurber, 1968; Miller, Bauchner, Hocking, Fontes, Kaminski, & Brandt, 1981; Miller & Stiff, 1993). The *information overload hypothesis* maintains that processing all incoming information would cause a cognitive overload in observers, who therefore would block out or overlook important cues (Bauchner, Kaplan, & Miller, 1980; Miller et al., 1981; Stiff, Miller, Sleight, Mongeau, Garlick, & Rogan, 1989). Both of these hypotheses, posited to account for the poorer accuracy rates attained when visual cues are present as compared to those situations where they are absent, predict that observers do not process the verbal information. However, Stiff et al. (1989) found that verbal information was processed by observers, although it was not used to make credibility judgments. The authors found partial support for an alternative hypothesis: the *situational familiarity hypothesis*, which maintains that observers in familiar situations use verbal

[9] Overall accuracy, that is, accuracy collapsing across the truthful and the deceptive statement, was always close to chance probability. Chi-square analyses on the hit / miss frequencies in neither case yielded statistically significant results; for the First Presentation: χ^2 (1) = 0.32, p = .571, for Moment-1-and-2 Judgments; χ^2 (1) = 0.08, p = .776, for Moment-3 Judgments; for the Second Presentation: χ^2 (1) = 1.63, p = .201, for Moment-1-and-2 Judgments; χ^2 (1) = 1.20, p = .274, for Moment-3 Judgments. Similarly, in neither of the loglinear analyses that examined the relations among Moment, Value of Truth, and the Hit/Miss variable, was the Moment X Hit/Miss second-order effect significant. These results lend further support to Levine, Park, and McCornack's (1999) arguments in favour of examining, in the field of the detection of deception, the separate accuracy for truthful and deceptive communications instead of focusing on the overall accuracy rate.

[10] It is important to keep in mind that this lack of significance in Experiment 1 was also apparent for S1T, which was the truthful account we used in the present study.

cues, since they can "visualize" the situation and assess the validity of verbal information (systematic processing), while observers in unfamiliar situations use, to some extent, nonverbal information, because there is little basis for evaluating the verbal content (heuristic processing). The observers we used were unfamiliar with the situation. In Experiment 1 the visual information was available; thus, they may have relied too much on that kind of information as a basis for their judgments.

Results from further explorations on the data of Experiment 1 support this explanation: when asked to indicate the cues they had used to make their judgments, observers reported significantly more nonverbal indicators, especially visual ones, than verbal cues (Masip, Garrido, & Rojas-Díaz, 2001; see also Garrido, Masip, Herrero, & Rojas-Díaz, 2000). This attention devoted exclusively to visual indicators may have caused accuracy for deceptive statements to decrease over time (as an increasing number of misleading visual indicators were being shown), while accuracy at judging the truthful statements did not rise in Experiment 1, since the most revealing cues are verbal, that is, just those cues observers did not attend to in that experiment. However, dynamic visual information, more misleading than the verbal one, was absent in Experiment 2. This may have led observers to pay close attention to the verbal cues, as well as to process those cues. This, in turn, may have contributed to the increased accuracy at judging the truthful statement at Moment 3.

This explanation should nevertheless be taken with caution. First, since verbal information is less misleading than the visible, hiding the latter should not only have resulted in an increase in accuracy over time for the truthful statement, but in addition it should have restricted the decrease in accuracy for the deceptive account. However, that decrease was significant not only in Experiment 1, but in Experiment 2 as well. A possible reason for that is that, after all, our sender was able to successfully control her verbal behavior. Second, despite experimental results showing the superiority of verbal information, as compared with the nonverbal, when it comes to making veracity judgments, our own research shows verbal cues may be processed in a biased manner. This, in turn, may have an effect upon the credibility judgments. For instance, police officers who participated as observers in Experiment 1 said the statements were implausible and contained verbal contradictions, while undergraduate students said they were plausible and verbally consistent. That is to say, each group of observers mentioned verbal indicators that were opposite to those mentioned by the other group, despite the fact that they all had been shown exactly the same videotapes. As a result of these perceptions the officers' tendency to judge the statements as deceptive was stronger than the undergraduates', and the latter's tendency to judge them as truthful was stronger than among officers. Similar results were found for a few nonverbal indicators (Garrido et al., 2000; Masip et al., 2001). Third, it would be inadequate to generalize from this second experiment (which was quite modest –its only pretension was to clarify some results found in Experiment 1–, where only two statements, both of them based on the same sequence, both of which were enacted by the same sender, were used) to other statements, witnesses, and situations. Caution is therefore strongly warranted when interpreting the results reported here, at least until further research replicates them.

Witnesses' Facial Appearance

In the discussion of Experiment 1 it was suggested that the lie bias, which was particularly strong at the beginning of the statement, could be caused by the sender's facial appearance. Therefore, in the present experiment several different faces were shown, to examine whether the lie bias of Experiment 1 or its time variation were influenced by the witness's appearance. However, contrary to our predictions, the senders' facial appearance had no influence either upon the overall lie bias, or upon its reduction over time. Therefore, these effects do not depend on the witness's facial appearance, at least, not for the range of faces used in this study. They are not

influenced by the witness' visible behavior (gestures and body movements) either, because that behavior was not shown in this experiment. Thus, they must be caused by verbal and paralinguistic cues, which were available in both studies.

The lack of influence of the face may nevertheless be due to several reasons. First, this was a competitive situation where observers were somewhat challenged to spot senders' lies. This differs from the cooperative interactions the average citizen is involved in his/her daily life, where truth is taken for granted and there is no motive to suspect the other is being deceptive. The nature of the task (detecting deception) may have raised observers' "state" suspicion (Levine & McCornack, 1991), making lie judgments more likely (Burgoon, Buller, Ebesu, & Rockwell, 1994; Stiff et al., 1992; Toris & DePaulo, 1984; Zuckerman, Driver, & Guadagno, 1985, p. 165), regardless of other factors such as the witnesses' facial appearance. Second, observers in this experiment were police officers. Garrido et al. (1997, 2003) and Sanderson (1978, cited by Bull, 1989) found a lie bias in officers' credibility judgments. Officers were also more accurate at judging lies than truths in recent studies conducted by Ekman, O'Sullivan and Frank (1999) and Porter, Woodworth and Birt (2000). Burgoon et al.'s (1994) military experts displayed a similar bias. It was suggested earlier that experts may hold a generalized communication suspicion (Levine & McCornack, 1991; see also Burgoon et al., 1994), which could increase their lie judgments. For instance, O'Sullivan, Ekman, and Friesen (1988) stated that "observers with a deception bias, because of their professional experience, for example as police officers or lawyers, may be more likely to view all behavior as deceptive and therefore have a heuristic which will permit them to classify deceptive behavior correctly, but which will be misleading in evaluating honest behavior" (p. 214). In addition, as suggested above, it appears reasonable that a lie bias heuristic could emerge for police officers, in the same way a truth bias cognitive heuristic emerges, according to Stiff et al. (1992), for relational partners. It may be that the strong lie bias displayed by our officers prevented them from being influenced by the subtle differences that existed between our senders. Perhaps students, whose lie bias was weaker, would have been sensitive to changes in the senders' facial appearance. Third, it may be that the differences in facial appearance between our senders were too small to have an influence upon credibility judgments. Despite the fact that observers' perceptions of their ages and physical attractiveness differed from one face to one another, all faces were perceived as in their twenties. Perhaps if a child's face, the face of a young person, that of a mature one, and an elderly person's face had been used very different results would have emerged. Also, the attractiveness of all three senders was close to average, ranging from 3.98 (face B) to 4.76 (face A), a rather small range on a 1-to-7-point scale. And, in addition, it is not only physical attractiveness that influences social judgments, but also mistaken identities, animal analogies, sickness similarities, babyfacedness, etc. (Zebrowitz, 1997). For example, recent research shows that, controlling for attractiveness, age and babyfacedness influence attributions of a series of traits and behavioral tendencies, including truthfulness / deceptiveness (Masip, Garrido, & Herrero, 2003a). Also, these facial characteristics have been found to influence the credibility judgments of written statements (Masip, Garrido, & Herrero, 2003b).

Finally, it could be argued that maybe participants did not pay any attention to the faces being shown, perhaps because they were fully aware that a still face with a neutral expression provides little information on whether the sender is lying or telling the truth. However it is unlikely that participants did not attend to the faces, because although facial appearance had no effect upon credibility judgments, they influenced the moment at which the decision was made. Regardless of the statement being judged (the truthful one or the deceptive one), when face A was being shown decisions were made at moments 1 or 2, and when face B was being shown decisions were made at the final moment. No clear tendencies emerged for Face C. Faces A and B differed from each

other both in terms of the age observers perceived in them and in attractiveness. Therefore, it appears that any of these two tendencies could account for the moment differences between Faces A and B. However, the perceived age of Face C did not differ significantly from that of Face A, and, just as Face A, it was perceived to be younger than Face B. Something similar happened with reference to attractiveness: Faces A and C did not differ from each other in this characteristic, and they both were judged as significantly more attractive than Face B. Therefore, if differences found between Faces A and B were due to age or attractiveness, Face C would have lined up with Face A, and this did not happen (actually, its non-significant tendencies were in the same direction as those of Face B). Therefore, the differences between the faces in terms of the moment the decision was made were not due to age or attractiveness, but, rather, to some other facial characteristic that was not taken into account in the present experiment.

General Discussion

Officers Versus Non-Officers

Contrary to our first prediction, police officers tended to make their decision as to whether the sender was truthful or deceptive later than non-officers. Although this effect was non-significant, it is possible that officers, knowing that collecting information from witnesses is important, paid attention to the sender's behavior for a longer time than undergraduates, thus deciding later than the students. However, their pronounced lie bias (Garrido et al., 1997, 2003), suggests that they were incapable of either collecting the right information or using it correctly to make accurate credibility judgments. In any case, these results do not discard the hypothesis that an a priori belief that the sender is deceptive can have an effect on officers' judgments. Indeed, it may be the case that police officers, instead of failing to process the incoming behavioral information, as they would do if, as suggested before, a cognitive heuristic such as that identified by Stiff et al. (1992) were in operation, actually do process that information, but in a biased manner aimed at finding support for their prior conceptions that the sender is being untruthful (conceptions based, for instance, on a GCS). In that case, officers would be unwittingly subjected to confirmation bias: "the tendency to interpret, seek, and create information in ways that verify existing beliefs" (Brehm & Kassin, 1993, p. 129). Recent, still unpublished research lends support to this idea (Masip, 2002; Masip et al., 2001).

Accuracy Over Time

In the studies reported here, a decrease in observers' accuracy at detecting deceptive statements was found as time went by, coupled with a similar increase in detecting truthful accounts, particularly when dynamic visual cues were not available to observers. This is probably due to the greater amount of misleading information in false performances as time goes by, and the greater amount of accurate information in the truthful accounts. Those time variations question the generalizability of findings of previous research, since most experiments on nonverbal detection of deception have been conducted using very small behavioral samples. It is apparent that receivers' detection accuracy depends on the moment in a long statement at which they make their decision, and it interacts with the value of truth of the statement: the moment truthful accounts are best detected is the same moment at which deceptive accounts are least detected, while overall accuracy is close to chance probability at any point in time.

Visual Versus Verbal Information

Our data seem to indicate that visual information prevented observers from properly using the growing verbal information that was presented in truthful statements. This is consistent with the distraction hypothesis, the information overload hypothesis, and the situational familiarity hypothesis, as well as with previous research showing the relative usefulness of verbal cues, compared to nonverbal ones, in judging credibility. Both when dynamic visual information was shown and when it was not available observers' accuracy at detecting deceptive accounts decreased as time went by. This suggests that, although extant research has shown that nonverbal cues are more misleading

than verbal ones, it seems that the audio channel (which conveys verbal and vocal information) can be successfully controlled by the sender in order to create a plausible lie and to appear honest. In fact, Ekman (1981) hypothesized that the verbal content would be very amenable to control, although meta-analyses show that the verbal information is very useful for making accurate credibility judgments: "the power (i.e., the accuracy) of the word, either written or spoken" (Zuckerman et al., 1981, p. 27).

Facial Appearance And Credibility Judgments

An accuracy level close to chance probability was expected for initial judgments in Experiment 1. However, a strong tendency to judge statements as deceptive was apparent for these early judgments. This tendency decreased over time. Since the only information available at the beginning of a statement is the sender's physical appearance, it was suggested that our witness's facial appearance could be responsible for that initial lie bias or its variation over time. However, we found no evidence of a face effect in the form of a demeanor bias in Experiment 2. Neither the existence of a lie bias nor its decrease over time was affected by the sender's facial appearance. However, the three faces that were used in Experiment 2 fell within the same age range and were close to average attractiveness, despite the significant differences that were found in the observers' attractiveness and age ratings. Research on how facial stereotypes may influence credibility judgments must be conducted. Recently, we completed a series of three experiments addressing this issue (Masip & Garrido, 2001b; Masip, et al., 2003a,b).

Caveats And Further Research

It must be acknowledged that the two studies reported here suffer from several methodological disadvantages. Aside from the problem of having used facial stimuli with rather small variations in the relevant facial features (age and babyfacedness indicators), three points must be mentioned here. First: those observers who took their decision at a given moment were not the same as those who decided at any other moment. This raises a question: were the differences found over time due to the influence of the moment variable or were they due to the impact of differences between the respondents who decided at different times? In addition, observers were not randomly assigned to the different moments, but they were free to make their decision at the time they preferred. Then, can we confidently assert that there is a strong initial lie bias or, rather, what happens is that those observers with the strongest lie bias decide at Moment 1 or Moment 2? This is unlikely, since officers, who were the most biased in Garrido et al.'s (1997, 2003) study, did not tend to decide early, but were the most biased at any moment in time. Certainly, that issue deserves further exploration. Second: As pointed out in the discussion of Experiment 1, the distinction between the different moments was a subjective one. Research on the influence of time on credibility judgments should use clear markers to differentiate between the moments of interest. Third: using only one sender is inappropriate. Several faces were used in Experiment 2, but in both experiments the speech was of the same person. The time profile found in both studies could be due to the verbal and/or paralinguistic idiosyncrasies of that sender. Thus, caution is warranted before generalizing these results to other senders.

In view of these problems the authors are about to conclude a study where a relatively large sample of senders (both males and females) watched videotapes depicting a theft. Later on, they were interviewed twice about the facts they had witnessed. In one case they had to tell the truth, in the other case they had to lie. Each interview had three questions. The answer to the first question was regarded as Moment 1, the answer to the second question as Moment 2, and the answer to the last question was taken as Moment 3. Witnesses' videotaped responses were shown to observers who had to judge the credibility of each statement three times: after watching the first answer, after watching the second one, and after watching the third (definitive judgment). This design overcomes the problems of the two experiments described in this report. Indeed,

its results will shed further light on the effect of the moment observers make their decision on credibility judgments and accuracy.

References

Alley, T. R. (1988). Physiognomy and social perception. In T. R. Alley (Ed.), *Social and applied aspects of perceiving faces* (pp. 167-186). Hillsdale, NJ: Lawrence Erlbaum Associates.

Alley, T. R., & Hildebrandt, K. A. (1988). Determinants and consequences of facial aesthetics. In T. R. Alley (Ed.), Social and applied aspects of perceiving faces (pp. 167-186). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bauchner, J. E., Kaplan, E. A., & Miller, G. R. (1980). Detecting deception: The relationship of available information to judgmental accuracy in initial encounters. *Human Communication Research., 6*(3), 253-264.

Becerra, A., Sánchez, F., & Carrera, P. (1989). Indicadores aislados versus patrón general expresivo en la detección de la mentira. *Estudios de Psicología, 38*(1), 21-29.

Berry, D. S., & Brownlow, S. (1989). Were the physiognomists right? Personality correlates of facial babyishness. *Personality and Social Psychology Bulletin, 15*(2), 266-279.

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology, 48*(2), 312-323.

Berry, D. S., & Zebrowitz, L. A. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin, 100*(1), 3-18.

Bond, C. F. Jr., Kahler, K. N., & Paolicelli, L. M. (1985). The miscommunication of deception: An adaptive perspective. *Journal of Experimental Social Psychology, 21*, 331-345.

Bond, C. F. Jr., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behavior, 12*(4), 295-307.

Brehm, S. S., & Kassin, S. M. (1993). *Social psychology. Second edition*. Boston: Houghton Mifflin Company.

Bruce, V., & Young, A. (1998). *In the eye of the beholder. The science of face perception*. Oxford: Oxford University Press.

Bull, R. (1982). Physical appearance and criminality. *Current Psychological Reviews, 2*, 269-282.

Bull, R. (1989). Can training enhance the detection of deception? In J. C. Yuille (Ed.), *Credibility assessment.* (pp. 83-99). Dordrecht: Kluwer Academic Publishers.

Bull, R., & McAlpine, S. (1998). Facial appearance and criminality. In A. Memon, A. Vrij, & R. Bull (Eds.), *Psychology and law. Truthfulness, accuracy and credibility* (pp. 59-76). London: McGraw-Hill.

Bull, R., & Rumsey, N. (1988). *The social psychology of facial appearance*. New York: Springer-Verlag.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203-242.

Buller, D. B., & Burgoon, J. K. (1998). Emotional expression in the deception process. In P. A. Andersen., & L. Guerrero. (Eds.), *Handbook of communication and emotion. Research, theory, applications and contexts* (pp. 381-402). San Diego.: Academic Press.

Burgoon, J. K., Buller, D. B., Ebesu, A. S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection. *Communication Monographs, 61*, 303-325.

DePaulo, B. M. (1991). Nonverbal behavior and self-presentation: A developmental perspective. In R. S. Feldman, & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 351-397). Cambridge: Camnbridge University Press.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin, 111*(2), 203-243.

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effect in the communication of deception. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 51-70). Dordrecht: Kluwer Academic Publishers.

DePaulo, B. M., & Pfeiffer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology, 16*(3), 249-267.

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology, 37*(10), 1713-1722.

DePaulo, B. M., Stone, J. I., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schlenker (Ed.), *The self and social life* (pp. 323-370). New York: McGraw-Hill.

DePaulo, B. M., Zuckerman, M., & Rosenthal, R. (1980a). Detecting deception. Modality effects. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 125-162). London: Sage.

DePaulo, B. M., Zuckerman, M., & Rosenthal, R. (1980b). Humans as lie detectors. *Journal of Communication, 30*, 129-139.

DePaulo, P. J., DePaulo, B. M., Tang, J., & Swaim, G. W. (1989). Lying and detecting lies in organizations. In R. A. Giacalone, & P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 377-393.). Hillsdale, NJ: Lawrence Erlbaum.

Ekman, P. (1981). Mistakes when deceiving. *Annals of the New York Academy of Sciences, 364*, 269-278.

Ekman, P. (1989). Why lies fail and what behaviors betray a lie. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 71-81). Dordrecht: Kluwer Academic Publishers.

Ekman, P. (1992). *Telling lies. Clues to deceit in the marketplace, politics, and marriage.* (2nd. ed.). New York: W. W. Norton & Company.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32*, 88-106.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology, 29*(3), 288-298.

Ekman, P., & O'Sullivan, M. (1989). Hazards in lie detection. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 253-280). New York: Springer.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*(9), 913-920.

Ekman, P., O'Sullivan, M., & Frank, M. (1999). A few can catch a liar. *Psychological Science, 10*(3), 263-266.

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

Garrido, E., & Masip, J. (1999). How good are police officers at spotting lies? *Forensic Update, 58*, 14-21.

Garrido, E., & Masip, J. (2001). Previous exposure to the sender's behavior and accuracy at judging credibility. In R. Roesch, R. R. Corrado, & R. Dempster (Eds.), *Psychology in the courts. International advances in knowledge* (pp. 271-287). London: Routledge.

Garrido, E., Masip, J., & Herrero, C. (2003). *Police officers' credibility judgments: Accuracy and estimated ability.* Submitted for review.

Garrido, E., Masip, J., Herrero, C., & Rojas-Díaz, M. (2000). La detección del engaño a partir de claves conductuales por agentes de policía. In A. Ovejero, M. V. Moral, & P. Vivas (Eds.), *Aplicaciones en psicología social* (pp. 97-105). Madrid: Biblioteca Nueva.

Garrido, E., Masip, J., Herrero, C., & Tabernero, C. (1997). *Policemen's ability to discern truth from deception of testimony.* Paper presented to the 7th European Conference of Psychology and Law, Stockholm, 3-6 September 1997.

Geis, F. L., & Moon, T. H. (1981). Machiavellianism and deception. *Journal of Personality and Social Psychology, 41*(4), 766-775.

Greene, J. O., O'Hair, H. D., Cody, M. J., & Yen, C. (1985). Planning and control of behavior during deception. *Human Communication Research, 11*, 335-364.

Henderson, J., & Hess, A. K. (1982). *Detecting deception: The effects of training and socialization levels on verbal and nonverbal cue utilization and detection accuracy.* Unpublished manuscript. Auburn University, Auburn, AL.

Kalbfleisch, P. J. (1985). *Accuracy in deception detection: A quantitative review.* Doctoral Dissertation: Michigan State University.

Kalbfleisch, P. J. (1992). Deceit, distrust and the social milieu: Application of deception research in a troubled world. *Journal of Applied Communication Research, 20*(3), 308-334.

Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements: Does it work? *Social Behaviour, 2*, 1-17.

Köhnken, G. (1989). Behavioral correlates of statement credibility: Theories, paradigms, and results. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system* (pp. 271-289). London: Springer-Verlag.

Kraut, R. (1980). Humans as lie detectors. *Journal of Communication, 30*, 209-216.

Kraut, R., & Poe, D. (1980). Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. *Journal of Personality and Social Psychology, 39*(5), 784-798.

Leekam, S. R. (1992). Believing and deceiving: Steps to becoming a good liar. In S. J. Ceci, M. D.Leichtman, & M. Putnick (Eds.), *Cognitive and social factors in early deception* (pp. 47-62). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Levine, T. R., & McCornack, S. A. (1991). The dark side of trust: Conceptualizing and measuring types of communicative suspicion. *Communication Quarterly, 39*, 325-339.

Maier, N. R. F., & Thurber, J. A. (1968). Accuracy of judgments of deception when an interview is watched, heard, and read. *Personnel Psychology, 21*, 23-30.

Martín, Q., Cabero, M. T., & Ardanuy, R. (1997). *Paquetes estadísticos SPSS 8.0. Bases teóricas. Prácticas propuestas, resueltas y comentadas*. Salamanca: Editorial Hespérides.

Masip, J. (2002). *La evaluación de la credibilidad del testimonio a partir de los indicadores conductuales en el contexto jurídico penal*. Unpublished Doctoral Dissertation. Department of Social Psychology and Anthropology. University of Salamanca.

Masip, J., & Garrido, E. (1999). Evaluación psicológica de la credibilidad: Contextualización teórica y paradigmas evaluativos. In A. P. Soares, S. Araujo, & S. Caires (Eds), *Avaliação psicológica: Formas e contextos* (Vol. VI, pp. 504-526). Braga: Associação dos Psicólogos Portugueses (APPORT).

Masip, J., & Garrido, E. (2000). La evaluación de la credibilidad del testimonio en contextos judiciales a partir de indicadores conductuales. *Anuario de Psicología Jurídica*, *10*, 93-131.

Masip, J., & Garrido, E. (2001a). La evaluación psicológica de la credibilidad del testimonio. In F. Jiménez (Ed.), *Evaluación psicológica forense, 1. Fuentes de información, abusos sexuales, testimonio, peligrosidad y reincidencia* (pp. 141-204). Salamanca: Amarú.

Masip, J., & Garrido, E. (2001b). Is there a kernel of truth in judgements of deceptiveness? *Anales de Psicología, 17*(1), 101-120. (Available online at: http://www.um.es/facpsi/analesps/v17/v17_1/08-17_1.pdf).

Masip, J., Garrido, E., & Herrero, C. (2003a). *Facial appearance and impressions of credibility: The effects of facial babyishness and age on person perception.* Submitted for review.

Masip, J., Garrido, E., & Herrero, C. (2003b). *Facial appearance and judgments of credibility: The effects of facial babyishness and age on statement credibility.* Submitted for review.

Masip, J., Garrido, E., & Rojas-Díaz, M. (2001). B*eliefs about indicators of deception and truthfulness in a specific situation.* Paper presented to the 11th European Conference of Psychology and Law, Lisbon, 5-8 June 2001.

Miller, G. R., Bauchner, J. E., Hocking, J. E., Fontes, N. E., Kaminski, E. P., & Brandt, D. R. (1981). ...and nothing but the truth. How well can observers detect deceptive testimony? In B. D. Sales (Ed.), *Perspectives in law and psychology, vol. II: The trial process.* (pp. 145-179). New York.: Plenum Press.

Miller, G. R., & Burgoon, J. K. (1982). Factors affecting assessments of witness credibility. In N. Kerr, & R. Bray (Eds.), *The psychology of the courtroom* (pp. 169-194). New York: Academic Press.

Miller, G. R., deTurck, M. A., & Kalbfleisch, P. J. (1983). Self-monitoring, rehearsal, and deceptive communication. *Human Communication Research, 10*, 97-107.

Miller, G. R., & Stiff, J. B. (1992). Applied issues in studying deceptive communication. In R. S. Feldman (Ed.), *Applications of nonverbal theories and research* (pp. 217-237). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Miller, G. R., & Stiff, J. B. (1993). *Deceptive communication*. Newbury Park: Sage.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology, 30*, 93-161.

O'Sullivan, M., Ekman, P., & Friesen, W. V. (1988). The effect of comparisons on detecting deceit. *Journal of Nonverbal Behavior, 12*(3), 203-215.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24*(6), 643-658.

Riggio, R. E., & Friedman, H. S. (1983). Individual differences and cues to deception. *Journal of Personality and Social Psychology, 45*(4), 899-915.

Riggio, R. E., Tucker, J., & Throckmorton, D. (1987). Social skills and deception ability. *Personality and Social Psychology Bulletin, 13*, 568-577.

Riggio, R. E., Tucker, J., & Widaman, K. F. (1987). Verbal and nonverbal cues as mediators of deception ability. *Journal of Nonverbal Behavior, 11*(3), 126-145.

Shepherd, J. (1989). The face and social attribution. In A. W. Young, & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 398-409). Amsterdam: North Holland.

Stiff, J. B., Kim, H. J., & Ramesh, C. N. (1992). Truth biases and aroused suspicion in relational deception. *Communication Research, 19*(3), 326-345.

Stiff, J. B., Miller, G. R., Sleight, C., Mongeau, P., Garlick, R., & Rogan, R. (1989). Explanations for visual cue primacy in judgments of honesty and deceit. *Journal of Personality and Social Psychology, 56*(4), 555-564.

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

Toris, C., & DePaulo, B. M. (1984). Effects of actual deception and suspiciouness of deception on interpersonal perceptions. *Journal of Personality and Social Psychology, 47*(5), 1063-1073.

Vrij, A. (1992). Credibility judgments of detectives: The impact of nonverbal behavior, social skills, and physical characteristics on impression formation. *The Journal of Social Psychology, 133*(5), 601-610.

Vrij, A. (1998). Nonverbal communication and credibility. In A. Memon, A. Vrij, & R. Bull. (Eds.), *Psychology and law. Truthfulness, accuracy and credibility* (pp. 32-58). New York: McGraw-Hill.

Vrij, A. (2000). *Detecting lies and deceit. The psychology of lying and the implications for professional practice.* Chichester, NJ: Wiley.

Vrij, A., Akehurst, L., & Morris, P. (1997). Individual differences in hand movements during deception. *Journal of Nonverbal Behavior, 21*(2), 87-102.

Vrij, A., & Graham, S. (1997). Individual differences between liars and the ability to detect lies. *Expert Evidence*, *5*(4), 144-148.

Vrij, A., & Winkel, F. W. (1993). Objective and subjective indicators of deception. *Issues in Criminological and Legal Psychology, 20*, 51-57.

Zebrowitz, L. A. (1997). *Reading faces. Window to the soul?* Boulder, Colorado: Westview Press.

Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology, 28*(6), 1143-1152.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). "Wide-eyed" and "crooked-faced": Determinants of perceived and real honesty across the life span. *Personality and Social Psychology Bulletin, 22*(12), 1258-1269.

Zuckerman, M., DeFrank, R. S., Hall, J. A., Larrace, D. T., & Rosenthal, R. (1979). Facial and vocal cues of deception and honesty. *Journal of Experimental Social Psychology, 15*, 378-396.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology, 14*, 1-59.

Zuckerman, M., & Driver, R. (1985). Telling lies: Verbal and nonverbal correlates of deception. In A. W. Siegman, & S. Feldstein (Eds.), *Multichannel integrations of nonverbal behaviors* (pp. 129-147). Hillsdale, NJ: Lawrence Erlbaum.

Zuckerman, M., Driver, R., & Guadagno, N. S. (1985). Effects of segmentation patterns on the perception of deception. *Journal of Nonverbal Behavior, 9*(3), 160-168.

APPENDIX 1: Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Occupation X Value Of Truth X Moment X Accuracy hierarchical loglinear analysis for *S1* in Experiment 1.

	Partial association chi-square	*p*	λ	*z*
First-order effect				
Moment	4.09	.043	Moments 1 and 2: .04; Moment 3: -.04	Moments 1 and 2: 0.55; Moment 3: -0.55
Second-order effects				
Moment X Hit/miss	12.25	.001	Mom. 1 & 2: Miss -.47, Hit .47; Moment 3: Miss .47, Hit -.47	Mom. 1 & 2: Miss -2.80, Hit 2.80; Moment 3: Miss 2.80, Hit -2.80
Value of truth X Hit/miss	49.63	.000	Deceptive: Miss -.47, Hit .47; Truthful: Miss .47, Hit -.47	Deceptive: Miss -6.20, Hit 6.20; Truthful: Miss 6.20, Hit -6.20
Occupation X Moment	3.24	.072	Officers: Moments 1 and 2 -.15, Moment 3 .15; Students: Moments 1 and 2 .15, Moment 3 -.15	Officers: Moments 1 and 2 -1.92, Moment 3 1.92; Students: Moments 1 and 2 1.92, Moment 3 -1.92
Third-order effects				
Value of truth X Moment X Hit/miss	11.40	.001	Deceptive: Moments 1 and 2 Miss -.24, Hit .24; Moment 3 Miss .24, Hit -.24. Truthful: Moments 1 and 2 Miss .24, Hit -.24; Moment 3 Miss -.24, Hit .24	Deceptive: Moments 1 and 2 Miss -3.12, Hit 3.12; Moment 3 Miss 3.12, Hit -3.12. Truthful: Moments 1 and 2 Miss 3.12, Hit -3.12; Moment 3 Miss -3.12, Hit 3.12
Occupation X Value of truth X Hit/miss	8.13	.004	Officers: Deceptive Miss -.20, Hit .20; Truthful Miss .20, Hit -.20. Students: Deceptive Miss .20, Hit -.20; Truthful Miss -.20, Hit .20	Officers: Deceptive Miss -2.66, Hit 2.66; Truthful Miss 2.66, Hit -2.66. Students: Deceptive Miss 2.66, Hit -2.66; Truthful Miss -2.66, Hit 2.66

Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Occupation X Value Of Truth X Moment X Accuracy hierarchical loglinear analysis for *S2* in Experiment 1.

	Partial association chi-square	*p*			λ				*z*			
First-order effects												
Hit/miss	1	.000			Miss	Hit			Miss	Hit		
					-.19	.19			-2.79	2.79		
Moment		.002			Moments 1 and 2	Moment 3			Moments 1 and 2	Moment 3		
					.13	-.13			1.96	-1.96		
Second-order effects												
Value of truth X Hit/miss		.049			Miss	Hit			Miss	Hit		
				Deceptive	-.10	.10		Deceptive	-1.47	1.47		
				Truthful	.10	-.10		Truthful	1.47	-1.47		
Value of truth X Moment	8	.001			Moments 1 and 2	Moment 3			Moments 1 and 2	Moment 3		
				Deceptive	.19	-.19		Deceptive	2.82	-2.82		
				Truthful	-.19	.19		Truthful	-2.82	2.82		
Third-order effects												
Value of truth X Moment X Hit/miss	5	.000				Miss	Hit				Miss	Hit
			Deceptive	Moments 1 and 2		-.24	.24	Deceptive	Moments 1 and 2		-3.58	3.58
				Moment 3		.24	-.24		Moment 3		3.58	-3.58
			Truthful	Moments 1 and 2		.24	-.24	Truthful	Moments 1 and 2		3.58	-3.58
				Moment 3		-.24	.24		Moment 3		-3.58	3.58
Occupation X Value of truth X Hit/miss		.008				Miss	Hit				Miss	Hit
			Officers	Deceptive		-.17	.17	Officers	Deceptive		-2.44	2.44
				Truthful		.17	-.17		Truthful		2.44	-2.44
			Students	Deceptive		.17	-.17	Students	Deceptive		2.44	-2.44
				Truthful		-.17	.17		Truthful		-2.44	2.44

APPENDIX 2: Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Judgment X Face X Moment hierarchical loglinear analysis for S1 *Truthful* in Experiment 2.

	Partial association chi-square	*p*	λ	*z*
First-order effect				
Judgment (of truthfulness / of deceptiveness)	38	.000	J. of Truthfulness: -.44; J. of Deceptiveness: .44	J. of Truthfulness: -5.58; J. of Deceptiveness: 5.58
Second-order effects				
Judgment X Moment	45	.000	Moments 1 and 2 / Moment 3: J.Truthfulness -.29 / .29; J.Deceptivnss .29 / -.29	Moments 1 and 2 / Moment 3: J.Truthfulness -3.70 / 3.70; J.Deceptivnss 3.70 / -3.70
Moment X Face	2	.017	Face A / Face B / Face C: Moments 1 & 2 .23 / -.18 / -.05; Moment 3 -.23 / .18 / .05	Face A / Face B / Face C: Moments 1 & 2 2.14 / -1.61 / -0.47; Moment 3 -2.14 / 1.61 / 0.47

Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Judgment X Face X Moment hierarchical loglinear analysis for S1 *Deceptive* in Experiment 2.

	Partial association chi-square	*p*	λ	*z*
First-order effect				
Judgment (of truthfulness / of deceptiveness)	52	.000	J. of Truthfulness: -.55; J. of Deceptiveness: .55	J. of Truthfulness: -6.46; J. of Deceptiveness: 6.46
Second-order effects				
Judgment X Moment	55	.001	J.Truthfulness: Moments 1 and 2: -.28, Moment 3: .28; J.Deceptivnss: Moments 1 and 2: .28, Moment 3: -.28	J.Truthfulness: Moments 1 and 2: -3.27, Moment 3: 3.27; J.Deceptivnss: Moments 1 and 2: 3.27, Moment 3: -3.27
Moment X Face	70	.000	Moments 1 & 2: Face A .44, Face B -.29, Face C -.15; Moment 3: Face A -.44, Face B .29, Face C .15	Moments 1 & 2: Face A 3.64, Face B -2.53, Face C -1.16; Moment 3: Face A -3.64, Face B 2.63, Face C 1.16

APPENDIX 3: Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Moment X Value of Truth X Hit/Miss hierarchical loglinear analysis for the *first* presentation in Experiment 2.

	Partial association chi-square	*p*	λ			z		
Second-order effect								
Value of truth X Hit/miss	29	.000		Miss	Hit		Miss	Hit
			Truthful Statemnt.	.55	-.55	Truthful Statemnt.	6 .70	-6.70
			Deceptive Sttmnt	-.55	.55	Deceptive Sttmnt	-6.70	6 .70
Third-order effect								
Moment X Value of truth X Hit/miss	54	.000			Miss	Hit		Miss	Hit
			Momnts 1 and 2 Moment 3	Truthful	.32	-.32	Momnts 1 and 2 Moment 3	Truthful	3.95	-3.95
				Deceptive	-.32	.32		Deceptive	-3.95	3.95
				Truthful	-.32	.32		Truthful	-3.95	3.95
				Deceptive	.32	-.32		Deceptive	3.95	-3.95

Partial association tests, parameter estimates (λ) and *z* statistics corresponding to the significant effects in the Moment X Value Of Truth X Hit/Miss hierarchical loglinear analysis for the *second* presentation in Experiment 2.

	Partial association chi-square	*p*	λ			z		
Second-order effect								
Value of truth X Hit/miss	57	.000		Miss	Hit		Miss	Hit
			Truthful Statemnt.	.42	-.42	Truthful Statemnt.	5 .60	-5.60
			Deceptive Sttmnt	-.42	.42	Deceptive Sttmnt	-5.60	5 .60
Third-order effect								
Moment X Value of truth X Hit/miss	33	.001			Miss	Hit		Miss	Hit
			Momnts 1 and 2 Moment 3	Truthful	.25	-.25	Momnts 1 and 2 Moment 3	Truthful	3.35	-3.35
				Deceptive	-.25	.25		Deceptive	-3.35	3.35
				Truthful	-.25	.25		Truthful	-3.35	3.35
				Deceptive	.25	-.25		Deceptive	3.35	-3.35

The Use of Law Enforcement Polygraph Tests with Juveniles

Ron A. Craig and Carla Molder

Abstract

Law enforcement polygraph examiners responded to a survey regarding their use of the polygraph with juveniles, types of test and special procedures used, and any perceived limitations in using the test with this population. The results indicate that polygraph tests are administered to juveniles in a variety of law enforcement contexts Many examiners express concern over testing juveniles below age twelve. A majority of the examiners make no modifications when testing juveniles. However, several reported specific limitations in using the polygraph with juveniles under a certain age. Of greatest concern regarding the use of the polygraph with juveniles was the potential limitation related to the development of cognitive abilities and sustained attention. While the polygraph is being used with juveniles, little research exists regarding its use with this population. There is a critical need to further explore the validity of the polygraph with this population.

While the use of the polygraph to detect deception is often not allowed in court, the practice of polygraphing adult suspects as part of a criminal investigation is commonplace (Goldzband, 1999). Most state law enforcement agencies have access to a polygraph examiner, and utilize them for a variety of purposes including clearing of suspects or as a prelude to the interrogation process (Honts & Perry, 1992). In addition, many courts require periodic polygraph tests as a condition for parole or as a component of treatment for sex offenders (Blasingame, 1998). The results of periodic polygraph testing of the offender hold weight in the decision to continue parole and treatment. Over the past 25 years, there has been a substantial body of research conducted to examine the validity of using physiological changes, as measured by the standard polygraph test, to detect deception in adults (Honts, Raskin, & Kircher, 2002). Research on the use of the polygraph has examined multiple techniques including the relevant/irrelevant test, guilty or concealed knowledge test (GKT), control questions test (CQT), and more recently the directed lie test (DLT). This research had included highly controlled laboratory simulations and detailed analysis of field data (for reviews see Honts et al., 1995; Kircher & Raskin, 1992; Raskin, 1986). However, the primary focus of this research has been on applying the polygraph to an adult population.

While the number of juveniles being investigated for serious crimes has declined since its peak in the early 1990s, a significant number of violent crimes involve juvenile suspects and garner significant public attention (Office of Juvenile Justice and Delinquency Prevention, 2000). As greater attention and resources are being focused on the issue of juvenile violence and juvenile offenders, it is reasonable to infer that law

[1] This article is reprinted with permission from the Journal of Credibility Assessment and Witness Psychology taken from Volume 4, Number 1, pages 63 - 74. Please refer to http://truth.boisestate.edu for copies for reproduction under the original copyright that follows:

enforcement may be interested in using the polygraph to detect deception in suspects from this population. Juveniles suspected of criminal activity could be asked to submit to a polygraph test, and then have the results of the test used against them during the interrogation process. It is also possible that passing the test may eliminate them from suspicion. In addition to potential investigative use, the polygraph is being utilized as a part of juvenile probation, particularly for sex offenses (Oregon Administrative Rules, 1995). Some juvenile suspects have taken a polygraph test and attempted to introduce the results in court to support their innocence, successfully in some cases (Adang, 1995) and unsuccessfully in others (*Commonwealth of Mass. v. A Juvenile*, 1974; South Carolina, *In the Interest of Robert R.*, 2000). The present study explores the use of the polygraph with juveniles, under what circumstances it has been used, and if polygraph examiners have identified any potential limitations. There are no set national guidelines for the use of the polygraph with juveniles nor is there a minimum testing age.

Steven Adang, a law enforcement polygraph examiner, identified two cases where juvenile polygraph results had been admitted in court (1995). Adang argues that competency of the juvenile is of primary concern, and that "Assuming that the requirements for competency are met, proper state of mind can be found for the polygraph examinee" (p. 262). Adang also surveyed six "seasoned" polygraph examiners and their use of the polygraph to detect juvenile deception. The examiners reported the youngest juvenile they had given a polygraph to was between the ages of 6 and 14 (Mean 11.5) and the number of juvenile polygraphs ranged from 4 to 300. Examiners reported a cutoff age for the polygraph from 6 to 16 (Mdn 12.5), and that attention span was the primary concern in administering a test with these juveniles. It is important to note that one of the examiners surveyed did ethically object to the use of the polygraph with juveniles, except for criminal investigations. Most issues identified by the examiners were also considered to be issues with adult examinees as well, and very few modifications in the polygraph test given to juveniles were reported. One examiner expressed concern over the use of control questions as being ineffective because they are not understood by the child or may not be a "probable lie for the minor."

To date, there is a remarkable absence of research regarding the use of the polygraph with juveniles, particularly those under 16 years-of-age. From the existing handful of studies addressing the issue (Abrams, 1975; Craig, 1997; Lieblich, 1971; Voronin, Konovalov, & Serikov, 1969) only Craig (1997) and Abrams (1975) have conducted a laboratory simulation of the polygraph consistent with the use of the polygraph in an investigative context. Voronin et al. (1969) used a card/number deception task with subjects from 6 to adult; using skin resistance (SR) to identify the memorized target. For the 6- to 7-year-olds, no targets were correctly identified, and for the 8- to 12-year-olds, only 12% were correctly identified; both were significantly lower than identification rates for the older populations. Lieblich (1971) administered an information detection task, similar to the GKT, to 3- to 4-year-old Israeli children. Skin resistance (measured as GSR) was the only physiological measure recorded during the test. Lieblich found that the detection rates, based on adult criteria, did not differ from chance.

In the Abrams (1975) study, 40 juveniles between the grades of 4 and 8, approximately 9 to 13 years of age, were subjected to a GKT regarding whether they had been given a pack of cherry-flavored Life Savers. If successful in deceiving the examiner they would be allowed to keep the candy. Those who did not receive any candy were told to respond truthfully to the question. Detection rates were averaged across two judges, with the lowest rates reported for those in the 4th and 5th grades (69% & 57% respectfully). Detection rates for older juveniles (6th, 7th, and 8th grade) were between 83% and 94%. Based on these findings, Abrams recommended caution when using the polygraph to detect deception in those under the age of 11. Abrams expressed particular concern regarding average intelligence of the juvenile, though he did not measure intelligence level in his study. In addition, Abrams failed to report whether the errors in detection were false positive or false negative. While there are limitations to the

Abrams study, including a small sample size, relative weak manipulation, and use of the GKT (a test not commonly used in the field), it does raise important questions and concerns regarding the effectiveness of using the polygraph with juveniles.

In Craig (1997), 9 to 15 years-olds participated in a mock crime scenario where the juveniles were accused of tearing a page from a book. Half of the subjects had torn the page out and were instructed to deny their involvement; the other half truthfully denied the act. All participants were given three tickets to a movie theater and were instructed that in order to keep the tickets they needed to convince the examiner they had not torn the book. Participants were given a Directed Lie Test (DLT) polygraph exam regarding the book. The DLT uses control questions that specifically instruct the participant to lie to them, compared to the probable lie Control Question Test (CQT), where the participant's lying to the question is generated through the manipulation in the pretest interview (Horowitz, Kircher, Honts, & Raskin, 1997). Thus, the DLT was selected, based on the researcher's concerns over the potential inability of younger children to meet the cognitive demand of the more common probable lie CQT. All participants were allowed to keep the tickets regardless of performance during the polygraph exam. Using the CPS scoring algorithm developed by Kircher and Raskin (1988) for scoring adult polygraph exams, 72.9% of the juveniles were correctly identified. This scoring method was more accurate at detecting innocent subjects (88.1%) than guilty (57.1 %). These error rates are inconsistent with the higher false positive rates found in adult studies (Horowitz et al. 1997). Craig (1997) produced a discriminant function based on the juvenile data for determining deception. The function, equally effective at detecting both the deceptive and truthful cases, correctly identified 73.8% of the participants.

With such a dearth of research on the topic and the potential for the active use of the polygraph with a juvenile population, there is a need for examination of the topic. There are a variety of potential issues that may impact the effectiveness of using a polygraph with juveniles. First, the test asks a series of questions that require a certain level of cognitive sophistication to be effective, a concern raised by Abrams (1975) and Craig (1997). Next, developmental changes in attention span and ability to remain still for a significant period of time may impact the validity of the polygraph test (Craig, 1997; Lieblich, 1971). In addition, there may be physiological differences between adolescents and adults that may alter the test. These changes include the way the responding organ functions (i.e. increased skin resistance reactivity in juveniles and reactivity of the cardiovascular system) and the neurological mechanisms that drive the physiological changes (for a review of developmental psychophysiology see Porges & Fox, 1986). Some researchers have even questioned if birth order might influence the ability of the polygraph to detect deception (Waid & Orne, 1982; Budnick, Love, & Wisniewski, 1983). One of the first steps in understanding the use of the polygraph with juveniles is to assess how often it is used with this population, the methods of testing that are most often used, and if any special alterations are being made. In addition, it is important to assess if those who are giving the exams have identified any limitations in using the polygraph with juveniles and if they have a minimum age for testing. The present study attempts to assess these questions related to the use of the polygraph with juveniles in an investigative setting, using information obtained via anonymous survey from law enforcement examiners across the United States.

Methods

Participants

A sample of 101 polygraph examiners was obtained as a result of sending 400 anonymous surveys to local and state law enforcement polygraph examiners. Survey recipients were identified via a membership mailing list obtained from the American Polygraph Association. Respondents included 93 males and 8 females all of whom reported working for either a local (77%), state (21%) or federal (1%) law enforcement agency. Respondents' ages ranged from 27 to 64 (M = 45.24, SD = 8.10), they had between 1 and 30 years experience conducting polygraph tests (M= 8.9, SD = 7.5), and had conducted between 4 and 750 polygraph examinations of

adults in the past year (Mdn = 82, Interquartile Range = 39 - 200). The majority (57.4%) reported having a bachelors or higher degree, with all but 3% having some college education. Most examiners reported having attended one of three polygraph training programs: Reid & Associates 22.8%, Backster School of Lie Detection 20.8%, and Argenbright International Institute 19.8 %. There were a number of other schools and programs mentioned but none with over 3% of the respondents having been trained there.

Materials

A survey was developed which requested demographic information such as the examiners' age, sex, and educational background, as well as formal training and years of experience. The survey asked the examiner to report the "Number of Juveniles (under 16 years of age) you have given a polygraph" including the age of the youngest juvenile tested and to identify the number of adults they had tested. Specifics were requested regarding what age they considered a juvenile too young to be tested and what limitations, if any, they perceived may be influential when testing a juvenile. In addition to the open-ended request for potential limitations, examiners were asked to rate on a 7-point scale the importance of 12 specific items in determining whether or not to conduct a polygraph with a particular juvenile. The questions from the questionnaire are shown in Table 1. An additional set of questions asked examiners to select the percentage of juvenile polygraphs they had conducted related to specific types of crimes from the following ordinal scale: none, 1%, 5%, 10%, 25%, 50%, 75%, 90+%. Types of crimes included property crime, drug crimes, murder, sexual assault, child sexual abuse, gang activity, as well polygraphs as a condition of probation or parole and polygraph of a juvenile witness.

Procedure

Members of the American Polygraph Association, identified via address as working in the legal system, were sent a package containing the anonymous survey and a self-addressed prepaid return envelope. The cover page of the survey identified researcher's affiliation, gave a brief explanation of purpose of the survey, addressed anonymity of responses, and requested the recipient participation. No follow-up requests or additional measures were taken to increase

Table 1 Questions From Survey Of Polygraph Examiners

In your opinion is there a cut-off age, where you feel a polygraph should not be given or is ineffective?

Age:

Why?

In making a determination on whether or not to polygraph a juvenile how important are each of the listed factors about the juvenile?

	Not Important			**Neutral**			**Very Important**
Attention span	☐	☐	☐	☐	☐	☐	☐
Ability to sit still for extended period	☐	☐	☐	☐	☐	☐	☐
The juvenile having ADHD	☐	☐	☐	☐	☐	☐	☐
Understanding of the truth vs. a lie	☐	☐	☐	☐	☐	☐	☐
Lower than average intelligence	☐	☐	☐	☐	☐	☐	☐
The juvenile having older siblings	☐	☐	☐	☐	☐	☐	☐
The juvenile having younger siblings	☐	☐	☐	☐	☐	☐	☐
Previous experience with a polygraph	☐	☐	☐	☐	☐	☐	☐
A history of telling lies	☐	☐	☐	☐	☐	☐	☐
Difficulty in school	☐	☐	☐	☐	☐	☐	☐
A history of aggressive behavior	☐	☐	☐	☐	☐	☐	☐
Having been abused	☐	☐	☐	☐	☐	☐	☐

response rate. To ensure confidentiality the surveys were numbered in the order by which they were returned.

Written responses to the question regarding the limitations in testing a juvenile below the minimum age listed by each examiner were coded to identify the specific type of limitation the examiner felt was significant. A coding system with 8 independent criteria was developed which categorized responses in separate domains. The domains coded for included cognitive limitations, moral development, training, and ethical concerns (Table 2). Two raters coded a total of 101 written statements for the presence of the 8 criteria resulting in a total of 140 coded remarks; a single statement could be coded with more than one criteria. The raters achieved an 80% agreement (Cohen's Kappa .765). Any differences in coding between the two raters were resolved through discussion.

Results

Analysis of the 101 law enforcement polygraph examiners responses indicates that 74.3% reported having tested at least one juvenile (under the age of 16), with those examiners having given between 1 and 1000 juveniles a polygraph test in their careers (Mdn = 6, IQR = 3 - 50). Figure 1 presents the age of the youngest juvenile examiners reported testing, M= 13.10, SD = 1.91. For those examiners who reported having given a polygraph to a juvenile, 66.7% reported making no special alterations in the test. Examiners reported an average of 78.8% (SD 24.5) of juvenile polygraph subjects were male. In response to the predominant ethnicity of juvenile subjects tested, examiners reported 81.5% White, 12.3% Black, 4.94% Hispanic, and 1.2% Asian. During the last year, examiners reported having tested significantly more adults (*Mdn* = 82) than juveniles (*Mdn* = 6, Wilcox test, W = 453, p <. 001). When

Figure 1. Age of the youngest juvenile ever tested by the respondents.

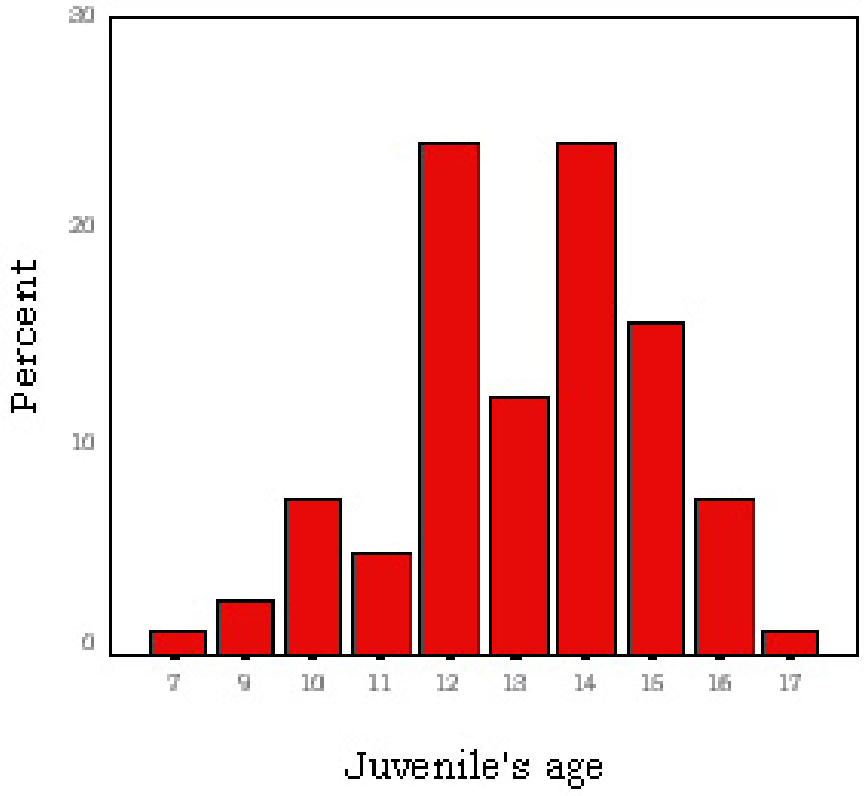


Figure 2. Reported minimum ages for polygraph tests

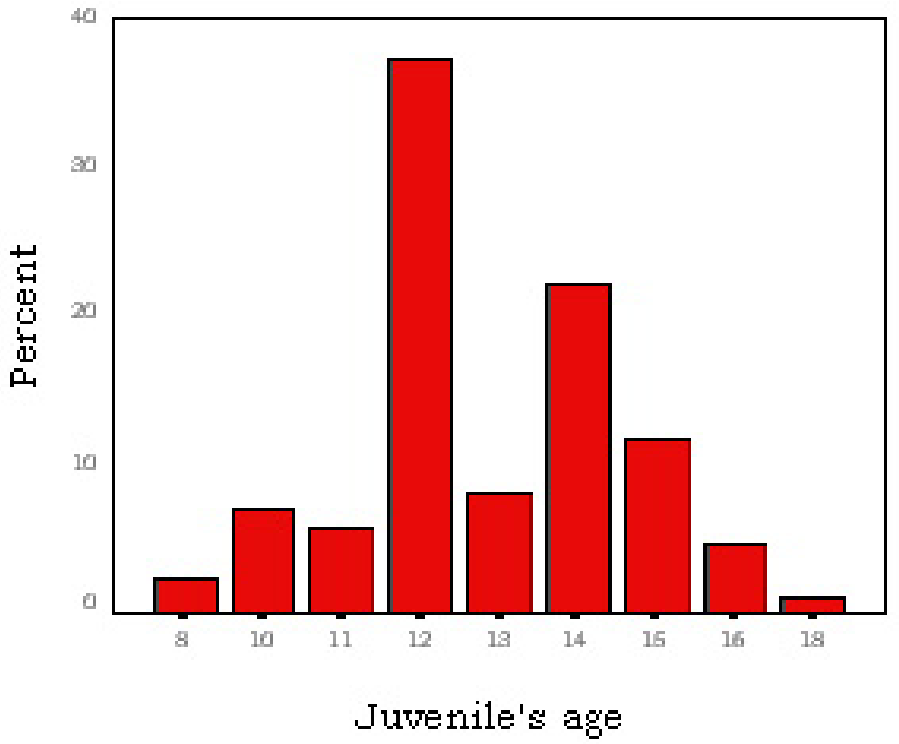


Table 2. Frequency of Reported Limitations

	Frequency	Percent
Cognitive limitations	43	30.7
Inability to comply with the requirements of the test	22	15.7
Moral development limitations	19	13.6
Regulations	10	7.1
Physical limitations	8	5.7
Trained not to test below that age	3	2.1
Ethical or personal prohibitions	2	1.4
No limitations reported	18	12.9
Other	15	10.7
Total	140	100.0

asked to list the type of test they have used with juveniles, 81.2% reported using the Control Questions Test, 7.9%, a Guilty Knowledge Test, 7.9%, a Relevant-Irrelevant test, and 3% a Directed Lie Test

Only 26.7% of the examiners reported doing any form of pre-screening to identify if the juvenile would not be a good subject for the polygraph. The mean reported minimum age for conducting a polygraph examination was 12.84, *SD* = 1.79. Figure 2 illustrates the minimum ages reported for actual examination.

For the 140 coded responses to the limitations in giving a polygraph to a juvenile below the minimum age the examiner specified, a significant difference was found for the types of limitations cited (χ^2 (8, N=140) = 79.85, p < .05). Examiners identified insufficient cognitive skills as the most common concern (30.7%) as to why the polygraph may be ineffective with a juvenile. Frequencies of the various concerns about juvenile are presented in Table 2. In response to an open ended request regarding any changes made in the control questions when testing juveniles, 35.8% adjusted language to age appropriate and relevant levels and 24.5% of those making alterations reported changing or eliminating the time bars. Other less frequently mentioned alterations included using directed lie controls or being sure of probable lie (15.1%), and more clarification of questions in the pre-test interview (7.5%).

A principle factor analysis using varimax rotation of importance ratings for the 12 specified limitations revealed three separate factors (Eigenvalues > 1.0) accounting for 67% of the variance in the data. A loading cut of .50 for inclusion of a variable in interpretation of a factor was used. The first factor was loaded with items related to either the juvenile having been abused or their ability to conform to societal expectation. The second factors captured cognitive/attention issues, and the third addressed presence of siblings (Table 3). The third factor was dropped from further analysis since examiners felt the juvenile having a sibling was unimportant regarding the polygraph tests (93% reporting Neutral to Not important). A composite score was produced for each factor by calculating the averaged importance ratings (1 unimportant to 7 very important) for each (Chronbach's alphas .90 & .81, respectively) and analyzed using paired sample t-test to identify which factor investigators felt was of the greatest concern when polygraphing a juvenile. Consistent with the analysis of the coded responses, the cognitive/attention factor (M= 5.93, SD = .09) was rated as significantly more important (t (92) = -11.93, p<.001) than moral understanding/behavior or whether they had been abused (M = 4.40, SD = .13).

Table 3. Factors For Examiner Responses To Potential Limitations Importance Ratings.

Factor 1		Rotated Factor Loading			Communalities
	Having been abused	.868			.779
	A history of aggressive behavior	.860			.839
	A history of telling lies	.824			.699
	Difficulty in school	.780			.735
	Understanding of the truth v. lie	.513			.423
Factor 2					
	Ability to sit still for extended period		.806		.651
	Attention span		.772		.628
	The juvenile having ADHD		.622		.534
	Lower than average intelligence		.602		.491
Factor 3					
	The juvenile having younger siblings			.957	.948
	The juvenile having older siblings			.948	.946
Eigenvalues		4.50	1.95	1.61	
% of Variance		37.50	16.20	13.40	

A Friedman test identified a significant difference (χ^2 (11, N=87) = 156.24, p < .001) in the responses regarding the types of crimes juveniles had been polygraphed fort. A post-hoc Bonferroni-Dunn analysis of the difference of the sums of ranks scores (Table 4) identified three crime types for which examiners most often conducted polygraph with juveniles: property crime (Mean Rank = 8.2), rape or sexual assault (Mean Rank = 8.1), and child sexual abuse (Mean Rank = 7.6). Polygraphs for both property crime and sexual assault were significantly more likely than seven of the other crime types: murder, robbery, drug use, selling drugs, gang activity, being a witness, and probation/parole. The reported occurrence of a juvenile polygraph regarding an accusation of child sexual abuse was significantly more likely than three of the other crime types: drug use, selling drugs and gang activity. There was no significant difference between property crime, sexual assault, and child sexual abuse.

Table 4. Significant Post-Hoc Comparisons Of Sums Of Ranks For Precipitating Reasons For Juvenile Polygraphs In A Law Enforcement Context.

Type of criminal activity suspected		Sum of Ranks	Sum of Ranks Difference
Property Crime		714.27	
	Murder or attempted murder/assault		178.35
	Robbery		166.17
	Selling drugs/drug trafficking		222.72
	Drug use		242.73
	Gang Activity		238.28
	Witness		187.05
	Condition of Probation/parole		201.84
Rape or sexual Assault		711.66	
	Murder or attempted murder/assault		175.74
	Robbery		163.56
	Selling drugs/drug trafficking		220.11
	Drug use		240.12
	Gang Activity		235.77
	Witness		184.44
	Condition of Probation/parole		199.23
Child sexual abuse		662.07	
	Selling drugs/drug trafficking		170.52
	Drug use		190.53
	Gang Activity		186.18

$CD_{F\ (n=87,\ k=12)} = 161.23$ $p < .05$

Discussion

The results indicate that law enforcement examiners are actively using the polygraph to detect deception in juveniles. While examiners test significantly more adults in a single year than the number juveniles they have tested in their careers, a substantial number of juveniles are being given polygraph tests. In addition, several examiners used the polygraph with early adolescent populations including juveniles as young as 7. It is important to note that more than half of the respondents do not use any special modifications when testing a juvenile, treating them exactly like an adult during the test. The alterations examiners did report making when using the polygraph with juveniles focused primarily on the issues of time bars or linguistic alteration in the control questions to make them more developmentally appropriate.

While most examiners do not make special modifications when using the polygraph with a juvenile, examiners did identify specific limitations for testing this population. Based on these perceived limitations, many examiners believed that a polygraph should not be used with anyone

below the age of 12. Some of the most frequent limitations cited were that juveniles lacked cognitive skills and moral understanding to produce meaningful physiological responses to the various polygraph questions. These perceived limitations correspond with the fact that the most commonly used polygraph test indicated was the Control Question Test, a rather cognitively sophisticated test. Research using less cognitively demanding tests procedures, like the Directed Lie Test or the Guilty Knowledge Test might be useful in addressing these perceived limitations.

Decisions as to whether or not to give a test, the type of test employed, and any modifications that might be made are often left to the discretion of the individual polygraph examiner. ASTM Standard Guide for PDD Examinations (ASTM, 2000) requires only that "The examiner shall ensure that the examinee is a fit subject for testing to the extent legally practicable. (p. 816)" The American Association of Police Polygraphists (AAPP, 2001) asserts that examiner has final authority regarding the validity of using the polygraph with a juvenile. Only three respondents reported a department policy on testing (one prohibited tests on juveniles, one the minimum age was 13, and the other the minimum age was 10). In the jurisdiction that did not allow testing of juveniles, the respondent noted that this was due to state law that required the presence of the parent during the test. Ultimately, there is no uniform set of guidelines as to how the polygraph should be used with juveniles, what specific factors an examiner should look for in making their determination, or a set minimum age.

Future research should focus on the types of tests commonly used with adults to determine if they are accurate when conducted with juveniles; potential limitations identified by these examiners should be addressed by such research. The criminal contexts under which juveniles are given polygraph tests are also of interest. With the dearth of empirical research, examiners are left without a firm foundation on which to advocate the use or nonuse of the polygraph with juveniles. Since there is a potential impact of both cognitive and physiological development on the legitimacy of using a polygraph with a juvenile population, examiners should approach testing juveniles with caution.

In recent years the notion of trying a juvenile as an adult has become increasingly common. Many states have amended their juvenile justice laws and have adopted adult criminal sanctions pertaining to certain crimes where the juvenile may be treated as an adult (Griffin, Torbet, & Szymanski, 1998). In addition, the use of the polygraph as a tool either in therapy or to monitor juvenile sex offenders is of concern. If there are developmental barriers that limit the effectiveness of the polygraph, then additional testing and measures may need to be used to ensure the test is being implemented appropriately. Working from a perspective that juveniles will perform the same as adults on a polygraph is, at this point, unsupported by research, thus the validity of such an assertion is uncertain. Ultimately, it may be that the polygraph is an effective tool in detecting juvenile deception, it may not be, or it may need to be altered to accommodate for developmental factors; only with more empirical and field research will the answer be known.

References

American Association of Police Polygraphists. (2001). *Bylaws of the American Association of Police Polygraphists.* Retrieved November 5, 2002 from http://www.policepolygraph.org/bylaws.pdf

Abrams, S. (1975). The validity of the polygraph technique with children. *Journal of Police Science and Administration, 3*, 310-311.

Adang, S. R. (1995). The use of the polygraph with children. *Polygraph, 24*, 259-274.

American Society for Testing and Materials. (2000). E 2065-00: Standard Guide for PDD Examinations. *Annual Book of ASTM Standards, 14.02 ,* 822-823.

Blasingame, G. D. (1998). Suggested Clinical Uses of Polygraph in Community-Based Sexual Offender Treatment Programs. *Sexual Abuse: A Journal of Research and Treatment, 10,* 37-45.

Bundnick, J. A., Love, K. G., and Wisniewski, L. (1983). Predictors of liar/nonliar status: Birth order, age, reason for polygraph investigation, and previous arrest. *Journal of Police Science & Administraion, 11,* 402-404.

Commonwealth v. A Juvenile, 365 Mass. 421, 313 N.E.2d 120 (1974),

Craig, R. A. (1997). The use of physiological measures to detect deception in a juvenile population: possible cognitive developmental influences (Doctoral Dissertation, University of Utah). *Dissertation Abstracts International, 58,* AAT 9812967.

Goldzband, M. G.(1999). Polygraphy revisited: U.S. v. Scheffer." *Journal of the American Academy of Psychiatry & the Law, 27,* 133-142.

Griffin, P., Torbet, P., and Szymanski, L. (1998). *Trying Juveniles as Adults in Criminal Court: An Analysis of State Transfer Provisions*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

Horowitz S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology, 34,* 108-115.

Honts, C. R.& Perry, M. V. (1992). Polygraph admissibility: Changes and challenges. *Law & Human Behavior, 16,* 357-379.

Kircher, J. C. & Raskin, D. C. (1992). Polygraph techniques: History, controversies, and prospects. In, P. Suedfeld, Peter & P. E. Tetlock, (Eds.), *Psychology and social policy* (pp. 295-308). Washington, DC: Hemisphere Publishing Corp.

Lieblich, I. (1971). Manipulation of contrast between differential GSRs in very young children. *Psychophysiology, 7,* 436-441.

Office of Juvenile Justice and Delinquency Prevention. (2000). *Challenging the Myths* (US Department of Justice, NCJ 178993). Washington DC: NCJRS.

Oregon Administrative Rules, OL Ch. 422, 416-460-0030, 1995

Porges, S. W., & Fox, N. A. (1986). Developmental psychophysiology. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 611-625). New York: The Gulford Press.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional, and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review, 1986*, 29-74

Raskin, D. C., Honts, C. R, & Kircher, J. C. (1995). 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.

South Carolina, *In the Interest of Robert R.*, Opinion No. 3165 (Ct. App., May 22, 2000).

Voronin, L. G., Konovalov, V. F., & Serikov, I. S. (1970). Interaction of conscious and unconscious trace processes in the nervous system. *Doklandy Akademii Nauk SSSR, 195,* 12237-1239.

Waid, W. M. and Orne, M. T. (1982). The physiological detection of deception. *American Scientist, 70*, 402-409.