Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

VOLUME 49 **2020** **NUMBER 2**

Contents

Addendum to the 2011 Meta-analytic Survey – the Utah Four-Question Test ("Raskin Technnique") / ESS 73
Editorial Staff

A Brief Comment on the Inhalation/Exhalation Ratios in Polygraph Scoring 79
Donald J. Krapohl

Deception Detection through Text Analysis 82
Aiman Shariff and Ritika Motwani

Electrodermal Responses: When is Bigger Really Better? 104
Donald J. Krapohl

Bigger is Better for Automated Scoring: Analysis of Minimum Constraints for RQ/CQ Ratios 110
Raymond Nelson

Using Virtual Reality to Improve Memory Recall and Detection of Deception in Forensic Interviews 121
Joyce Yan Ting Sam, Lin Qiu, and Ky Phong Mai

Accuracy Effects for ESS and Three-Position Scores of Federal ZCT Exams Using the Grand Total Rule with Traditional/Federal and Multinomial Cutscores 157
Raymond Nelson

Modification of the AFMGQT to Accommodate Single-Issue Screening: The British One-issue Screening Test 176
Donald J. Krapohl, Donald Grubin, Tim Benson and Bernard Morris

Timely Non-deceptive Sexual History Polygraph Examinations are Correlated with Completion of Treatment but Not Correlated with Sexual Recidivism 184
James E. Konopasek and Johneen Manno

Published Semi-annually

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

Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

Subscription information: *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* is published semi-annually by the American Polygraph Association. Editorial Address is Editor@polygraph.org. Subscription rates for 2020: One year $150.00. Change of address: APA National Office, P.O. Box 8037 Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

Instructions to Authors

Scope

The journal *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* publishes articles about the psychophysiological detection of deception, and related areas. Authors are invited to submit manuscripts of original research, literature reviews, legal briefs, theoretical papers, instructional pieces, case histories, book reviews, short reports, and similar works. Special topics will be considered on an individual basis. A minimum standard for acceptance is that the paper be of general interest to practitioners, instructors and researchers of polygraphy. From time to time there will be a call for papers on specific topics.

Manuscript Submission

Manuscripts must be in English, and may be submitted, along with a cover letter, on electronic media (MS Word). The cover letter should include a telephone number, and e-mail address. All manuscripts will be subject to a formal peer-review. Authors may submit their manuscripts as an e-mail attachment with the cover letter included in the body of the e-mail to:

Editor@polygraph.org

As a condition of publication, authors agree that all text, figures, or other content in the submitted manuscript is correctly cited, and that the work, all or in part, is not under consideration for publication elsewhere. Authors also agree to give reasonable access to their data to APA members upon written request.

Manuscript Organization and Style

All manuscripts must be complete, balanced, and accurate. Authors should follow guidelines in the *Publications Manual of the American Psychological Association*. The manual can be found in most public and university libraries, or it can be ordered from: American Psychological Association Publications, 1200 17th Street, N.W., Washington, DC 20036, USA. Writers may exercise some freedom of style, but they will be held to a standard of clarity, organization, and accuracy. Authors are responsible for assuring their work includes correct citations. Consistent with the ethical standards of the discipline, the American Polygraph Association considers quotation of another's work without proper citation a grievous offense. The standard for nomenclature shall be the *Terminology Reference for the Science of Psycho-physiological Detection of Deception* (2012) which is available from the national office of the American Polygraph Association. Legal case citations should follow the *West* system.

Manuscript Review

An Associate Editor will handle papers, and the author may, at the discretion of the Associate Editor, communicate directly with him or her. For all submissions, every effort will be made to provide the author a review within 4 weeks of receipt of manuscript. Articles submitted for publication are evaluated according to several criteria including significance of the contribution to the polygraph field, clarity, accuracy, and consistency.

Copyright

Authors submitting a paper to the American Polygraph Association (APA) do so with the understanding that the copyright for the paper will be assigned to the American Polygraph Association if the paper is accepted for publication. The APA, however, will not put any limitation on the personal freedom of the author(s) to use material contained in the paper in other works, and request for republication will be granted if the senior author approves.

Addendum to the 2011 Meta-analytic Survey –

the Utah Four-Question Test ("Raskin Technique") / ESS

Editorial Staff

APA (2011) published the report of meta-analytic survey of validated polygraph techniques in preparation for evolving standards of practice which require the use of validated techniques in field polygraph practice. Two important aspects of the design of that project were the specification of requirements for inclusion in the study and a definition of what is meant by term polygraph technique. For the purpose of that survey, a polygraph technique was defined as a defined question sequence together with an analysis method. This definition was premised on an assumption that the effectiveness of a polygraph technique is, in part, a function of the recording data for analysis and interpretation and also analysis of the recorded data through a valid and structured process.

Inclusion in the meta-analytic survey required published and replicated studies showing test sensitivity, specificity, false-positive and false negative rates, in addition to the publication of the means and variance of the sampling distributions. The requirement for publication and replication was premised on an assumption that all research samples are biased – they are an imperfect representation of the population – and the fact that sampling statistics, if randomly selected, will converge towards the unknown population parameters according to the central limit theorem (Kwak & Kim, 2017). Researchers in all areas of social science make use of this theorem to develop tests and measures for amorphous phenomena such as personality traits, intellectual functioning, academic achievement, height, weight ,or any population referenced phenomena of interest.

Although the Utah three-question test was included in the meta-analytic survey, the Utah four-question test – sometimes referred to as the "Raskin technique" – was not included due to an absence of published information specific to this format. The Utah four-question test is mentioned by APA (2011), in footnote #50 on page 248, for its structural similarity to the AFMGQT, with an advisement that information can be generalized for the two named formats.

One important difference between the Utah four-question format and the AFMGQT is that the latter is commonly interpreted using the subtotal-score-rule (SSR) whereas the Utah four-question format is commonly interpreted with the grand-total-rule (GTR) or two-stage-rule (TSR). [See Nelson (2018) for a survey of polygraph decision rules.] Underlying the selection of a polygraph decision rule is an assumption as to whether the relevant questions are independent or non-independent (dependent).

Independence, in the scientific context, requires that the questions have no shared source of variance through which factors that influence responses to any question could also influence responses to other questions. The AFMGQT is commonly used in polygraph screening contexts in which relevant questions are formulated to investigate an array of behavioral concerns in the absence of any known incident or allegation, and are commonly interpreted with an assumption of independent criterion variance (notwithstanding that the examinee's attention will remain a potential dependency or influencing factor within a multiple issue exam). For reasons both psychological and statistical (i.e., multiplicity) multiple issue exams cannot provide the same level of accuracy or precision as single issue exams. However, multiple issue exams are useful in polygraph screening programs.

In contrast, the Utah four-question format is used as an event-specific diagnostic polygraph format – used to investigate the verac-

ity of an examinee's statements regarding a known incident. Because all of the relevant questions involve a single known incident or allegation, an assumption of independence is unfounded. A convenient and useful aspect of the Utah four-question single issue test is that any combination of primary relevant and secondary relevant questions (i.e., weak relevant, evidence connecting, guilty knowledge, helping, planning, participating, and role descriptive questions) is permitted. That is, field examiners are free to use any type of relevant question that best suits the circumstances and needs of the investigation. As a matter of practice at least one of the relevant questions is commonly a primary relevant question that describes the examinee's behavioral involvement the topic of the investigation. This is, in many ways, similar to the formulation of questions for the AFMGQT versions 2. The core aspects of the sequence itself [CQ, RQ, RQ, CQ, RQ, RQ, CQ] is structurally identical for the Utah four-question test and the AFMGQT. Another similarity for the two formats is that all CQs and RQs are rotated in a random or pseudo-random manner for each iteration of the question sequence.

Regardless of the combination of primary and secondary relevant questions, the Utah four-question test is interpreted with an assumption of shared criterion variance among the RQs – the examinee was either involved or not involved in the allegation or incident under investigation. An assumption of non-independent criterion variance forms the basis for the use of the GTR or TSR, and simplifies the assumptions and requirements both psychologically and statistically. The examinee is not subject to divided attentional demands, because all relevant questions pertain to the same incident or allegation. Multiplicity effects are not a factor when using the GTR, and are reduced through the use of a statistical correction when using the TSR.

Another important aspect of the Utah four-question format is that the use of four relevant questions, instead of three or two, will mean that the test result will be based on more information compared to three-question and two-question test formats. As a general principle, use of more data leads to increased precision or accuracy of quantitative conclusions. This is related to the *law-of-large-numbers* (Dekking, 2005; Révész, 1968) which holds that the frequency of occurrence of a random event converges towards its probability as the number of trials increases. It is the reason that larger samples are preferred to smaller samples – they can, if randomly selected, more closely approximately an unknown population parameter.

With the Utah four question test, the amount of data for a polygraph test with four RQs and up to five iterations of the question sequence is more than three times that of a test with three iterations of two RQs. The result of this is an increased in both test sensitivity and specificity, with a corresponding reduction in inconclusive results and increase in overall precision Another result is that the Utah four-question test can be more robust, and less vulnerable, in the context of difficult test data.

Raskin, Honts, Nelson and Handler (2015) published the results of a Monte Carlo study on the Utah four-question test, including both ESS and seven-position scores. Seed data for the Monte Carlo were N=100 exams from the University of Utah. Results were statistically undifferentiatable for the two scoring methods. For ESS scores with the TSR using alpha = .05/.05 for deception and truth-telling the unweighted accuracy rate for five iterations of the question sequence was .949 with an unweighted inconclusive rate of .020.

Nelson (2018) published a second study on the Raskin technique using data from th DoDPI confirmed case archive. Examinations were a sample of N=30 confirmed field cases that were conducted using the AFMGQT format. This format was described earlier as structurally similar to the Utah four question test; all cases consisted of three iterations of a question sequence that included a combination of primary and secondary relevant questions. Scored were obtained using an automated version of the ESS-M, and results were classified as deceptive or truthful using the TSR with alpha=.05/.05 deception and truthtelling. Unweighted accuracy was .929 with an inconclusive rate of .033.

Figure 1 shows a mean and standard deviation plot of the scores of the sampling distributions of the included Utah Four-Question

(Raskin) Technique studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not statistically significant *[F (1,88) = 0.673, (p = .414)]*, nor was the main effect for sampling distribution *[F (1,88) = 0, (p = .993)]*. One-way ANOVAs showed no significant differences in the scores of the two studies for either the deceptive samples *[F (1,44) = 0.018, (p = 0.895)]* or truthful samples *[F (1,44) = 0.02, (p = 0.888)]*.

Figure 1. Mean and standard deviations for the scores from truthful and deceptive samples with the Utah four-question format.

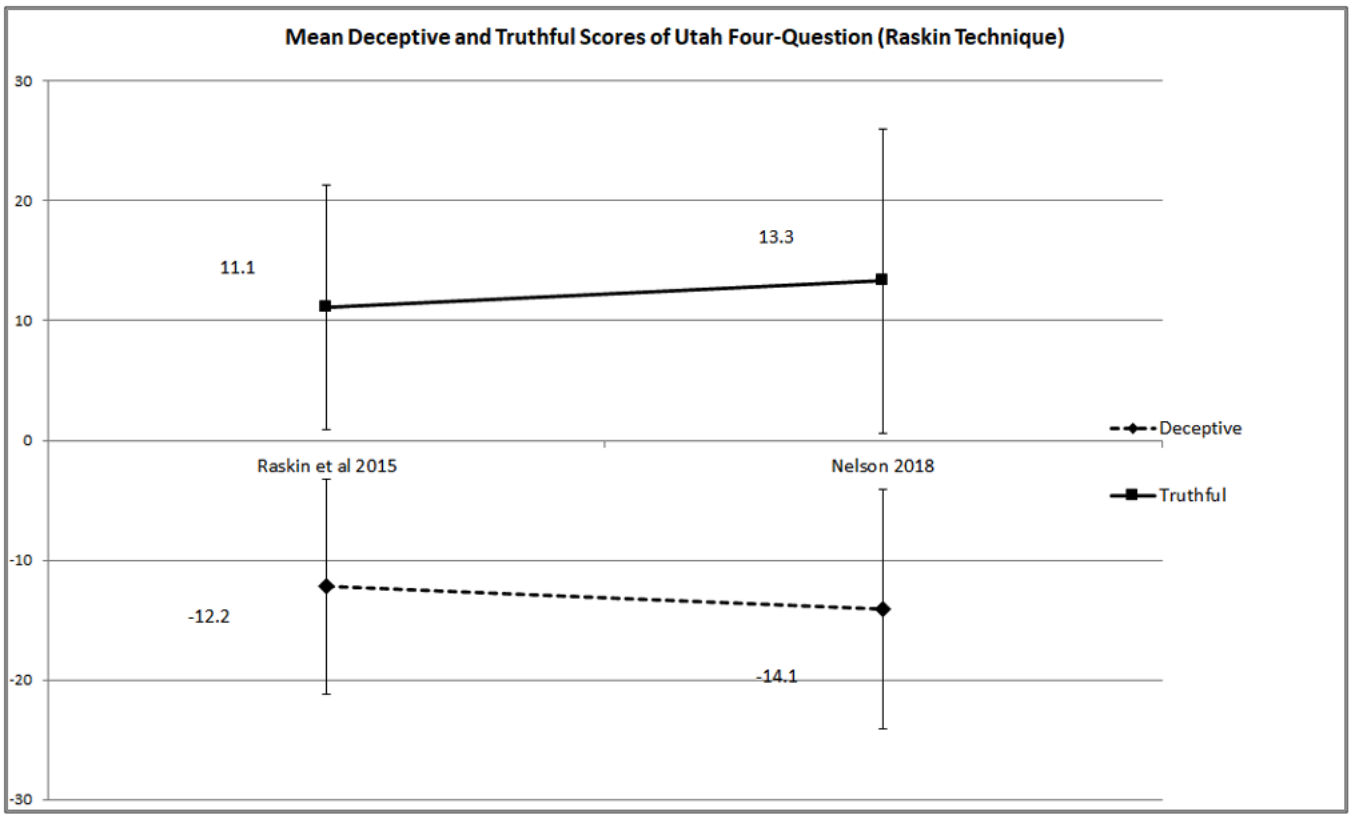


Table 1 shows the summary of the two combined on the Raskin technique.

Table 1. Summary of studies on the Raskin technique.	
Number of usable Studies	2
Total N	130
N Deceptive	65
N Truthful	65
Number of Examiners/Scorers	2
Total Scores	130
D Scores	65
T Scores	65
Mean D	-12.638
StDev D	10.154
Mean T	11.608
StDev T	9.854

Table 2 shows the profile and statistical confidence intervals for the criterion accuracy metrics.

Table 2. Criterion accuracy and confidence intervals for the Raskin technique.	
Unweighted Average Accuracy	.944 (.021) {.897 to .984}
Unweighted Inconclusives	.031 (.026) {.010 to .092}
Sensitivity	.923 (.033) {.852 to .984}
Specificity	.908 (.036) {.831 to .971}
FN Errors	.046 (.026) {.010 to .104}
FP Errors	.062 (.030) {.014 to .125}
D-INC	.031 (.021) {.010 to .078}
T-INC	.031 (.021) {.010 to .078}
PPV	.938 (.031) {.871 to .986}
NPV	.952 (.027) {.892 to .990}
D Correct	.952 (.027) {.893 to .990}
T Correct	.936 (.031) {.871 to .986}
Detection Efficiency Coefficient	.875 (.041) {.788 to .949}

Table 3 shows a summary of the individual studies.

Table 3. Summary of individual studies on the Raskin technique.		
Study	Raskin et al., (2015)	Nelson (2018)
Sample N	100	30
N Deceptive	50	15
N Truthful	50	15
Scorers	1	1
D Scores	50	15
T Scores	50	15
Total Scores	30	30
Mean D	-12.2	-14.1
StDev D	10.2	10.0
Mean T	11.1	13.3
StDev T	9.0	12.7
Unweighted Average Accuracy	.949	.929
Unweighted Inconclusives	.020	.067
Sensitivity	.940	.870
Specificity	.920	.870
FN Errors	.040	.067
FP Errors	.060	.067
D-INC	.020	.067
T-INC	.020	.067
PPV	.940	.929
NPV	.958	.929
D Correct	.959	.929
T Correct	.939	.929

The combined decision accuracy level of the Utah four-question test ("Raskin technique") studies, weighted for sample size and number of scorers, was .944 with a combined inconclusive rate of .031. The detection efficiency coefficient, calculated as the correlation between the categorical result coded as [+1, 0, -1] and the criterion state for each case coded as [+1, -1] was .875 with a 95% confidence interval from .788 to .949.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. [Electronic version] Retrieved September 15, 2020, from http://www .polygraph.org.

Dekking, Michel (2005). *A Modern Introduction to Probability and Statistics.* Springer.

Nelson, R. (2018). Credibility assessment using Bayesian credible intervals: a replication study of criterion accuracy using the ESS-M and event-specific polygraphs with four relevant questions. *Polygraph and Forensic Credibility Assessment*, 47 (1), 85-90.

Kwak, S. G. & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. Korean *Journal of Anesthesiology.* 70(2), 144–156.

Nelson, R. (2018b). Practical polygraph: a survey and description of decision rules. *APA Magazine,* 51(2), 127-133.

Raskin, D. C., Honts, C., Nelson, R. & Handler, M. (2015). Monte Carlo estimates of the validity of four-relevant-question polygraph examinations. *Polygraph 44(1)*, 1-27.

Révész, P. (1968). *The Laws of Large Numbers.* Academic Press.

A Brief Comment on the Inhalation/Exhalation Ratios in Polygraph Scoring

Donald J. Krapohl[1]

The first published attempt to use a systematic evaluation of breathing for deception detection was by Vittorio Benussi (1914). His method entailed the recording of 3 – 5 cycles of breathing immediately before and after his experimental subjects gave true or false spoken narratives regarding characters depicted on slips of paper. Respiration during speaking was ignored. The breathing cycles recorded before speaking normally occurred between looking at the image and beginning to describe it. Benussi reported remarkable accuracy for his method. He created an inspiration/expiration (I/E) ratio to represent the relative time taken by each of these two portions of the respiratory wave. Benussi's work was partially replicated by Burtt (1921) and Landis and Wiley (1926) who found a lower accuracy than did Benussi though the approach was still promising.

The way Benussi considered I/E ratios in deception detection would not be intuitive to most polygraph examiners. He wrote:

> "…in relation to the phase before the true statement expiration is slower in the subsequent phase, and in relation to the phase before the untrue statement expiration is more rapid in the subsequent phase. In other words: In regard to distribution between inspiration and expiration phase of a single breath the innervation of the respiratory muscles changes for lie and truth in opposite directions, in that in the latter in innervation of inspiration is relatively stronger in the phase preceding the statement than in the following phase, while in the former case it is weaker. These symptoms are so distinct that in many cases it would be completely sufficient to measure only two breaths, the one immediately preceding the statement and the only immediately following."

Said another way, truthtellers exhaled more slowly after speaking the truth than before speaking the truth (while contemplating the image). Liars showed the opposite trend. As is immediately apparent, this is not how the concept has been translated into polygraph practice.

Nonetheless, the recommended use of I/E ratios for scoring the pneumograph can be found in the recent polygraph literature and it is currently taught in some polygraph schools. The processes of Benussi, Burtt, Landis and Wiley are so dissimilar from polygraph testing it cannot be assumed their research bears directly on modern polygraphy. The question naturally arises as to what evidence there is for scoring changes in I/E ratios in the context of polygraph testing.

I was able to locate one study that directly evaluated the I/E ratio in polygraph testing. It was completed under US Government contract in 2005 (Kircher, Kristjansson, Gardner & Webb, 2005). The project was undertaken to assess 25 scoring features reported by Swinford (1999) as representing the Government's approach to polygraph scoring in that era. In that study they recruited 32 US government polygraph examiners to blindly score 80 con-

[1] The author is an APA Past President and regular contributor to this journal. Questions and comments can be sent to APAkrapohl@gmail.com.

The opinions expressed are solely those of the author, and do not necessarily represent the views of the Capital Center for Credibility Assessment (C3A) or the American Polygraph Association.

firmed polygraph cases, half of the cases being from deceptive examinees. Examiners in the study assigned seven-position scores to the 80 cases and separately documented the scoring features they used to make their numerical assignment. One of the assessed pneumograph tracing features was the change in the I/E ratio, both for two and for three respiration cycles. Separately, a computer algorithm evaluated the same cases.

With the advantage of our current understanding about polygraph scoring it might easily be predicted the Kircher team would find few of the Government's 25 scoring features were valid. Changes in the I/E ratio did not survive the cut. It was characterized in this very careful study as "unreliable and invalid." In addition, the use of I/E ratio changes by the 32 examiners was minimal and may signal a recognition the feature is not helpful.

There is a lack of evidence to support scoring I/E ratio changes in polygraph testing. Its inclusion in contemporary polygraph school lesson plans appears to be an uncorrected accident of history. Polygraph instructors and field examiners could profit from aligning with the current state of the evidence.

References

Benussi, V. (1914). *Die Atmungssymptome der Lüge*. In: *Archiv für Psychologie*. 31, S. 244–273. English translation appears in Polygraph (1975), vol 4(1), 52 – 76.

Burtt, H.E. (1921). The inspiration-expiration ratio during truth and falsehood. Journal of *Experiment Psychology, 4*(1), 1 – 23.

Kircher, J.C., Kristjansson, S.D., Gardner, M.K., and Webb, A. (2005). *Human and Computer Decision-making in the Psychophysiological Detection of Deception.* Final Report to the DoD Polygraph Institute. Project # DoDPI00 – P – 0002. Reprinted in Polygraph 41(2).

Landis, C., and Wiley, L.E. (1926). Changes in blood pressure and respiration during deception. *Journal of Comparative Psychology, 6*(1), 1–19.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph, 28*, 10-27.

Deception Detection through Text Analysis

Aiman Shariff[1] and Ritika Motwani[2]

Manipal Institute of Technology, Manipal[1]

Indian Institute of Information Technology Allahabad[2]

Abstract

Deception detection in humans is a long-standing issue; one that has immense significance across numerous aspects of life: from law enforcement, to business, to social media. In this paper, we seek to develop a textual approach to automated deception detection, and explore how machine learning and deep learning approaches can indicate deceptiveness in written and spoken text. We mainly analyze two sets of text data: online real and fake hotel reviews, and transcriptions of real life court trials. By including contextual features of text as well as psycho-linguistic features we are able to develop classifiers that perform well across both sets of text. Our machine learning classifier is able to achieve an accuracy of nearly 92% in the task of deceptive spam detection. We also demonstrate how a combination of textual and psycholinguistic features enables us to feed fewer features to our classifier without a significant drop in accuracy. Further, we observe the promise of deep learning techniques as our Convolutional Neural Network model achieves 90% accuracy in detecting deceptive spam. We also discuss the possibility of crafting malicious inputs to thwart our classifiers.

I. Introduction

The problem of deception detection is a long standing issue with great relevance to law enforcement, social media, and even everyday life. The most common methods for lie detection, such as polygraph tests, are based on physiological indicators. However, these techniques are not applicable to virtual deceit - online impostors and fake messages. Further, they require the physical presence of the subject. There have not been any significant advances towards creating an automated deception detection system, something that would solve a pressing need and deal with issues ranging from false testimonies in legal matters, to identity fraud on social media.

According to (Vrij, Edward, & Bull, 2000), there are three approaches to lie detection: non-verbal (observing an individual's gaze, movements etc.), verbal (analyzing the speech content), and physiological (heart rate, muscle activity etc.). A further work by Vrij et al. (2018), suggests that focusing on verbal, rather than non-verbal, cues helps better to distinguish between a liar and a truth teller.

In this paper, we aim to focus specifically on using linguistic techniques (in particular Natural Language Processing) to analyze deceptive language. Although acoustic (Mendels et al., 2017) and other non-linguistic features such as video (Abouelenien et al., 2017) are shown to be promising as well, we concentrate specifically on deceptive text as it is the most

[1] A note to the reader: the appendices at the end of this paper contain basic but useful explanations for the various mathematical and machine learning concepts that have been touched upon in this paper. Kindly refer as needed.

[2]For two matrices A and B of the same dimension m × n, the Hadamard product A B is a matrix of the same dimension as the operands, with elements given by (AB)ij=(Aij)(Bij).

commonly occurring medium of deception-relevant data on the web. Moreover, it shows strong potential as an indicator for deception (Krishnamurthy et al., 2018).[1]

II. Literature Review

Researchers have long been looking for reliable indicators of deception. Language-based techniques were spearheaded by Vrij, Edward, & Bull, (2000) who analyzed the usage of Criteria-Based Content-Analysis and Reality Monitoring as tools to predict deceptiveness in speech content. However, these human techniques can be subjective and automating them poses a large challenge. In 2003, De Paulo et al. performed a meta-analysis of around 120 cues for deception. Of these cues, a few were verbal in nature - such as word and phrase repetitions, while most of the cues pertained to the speaker's behavior - including eye shifts and facial expressions. Burgoon et al (2003) attempted to use cues to automate linguistic deception detection with individual cue analysis, as well as cluster analysis by C4.5 (a statistical classification algorithm), creating an early benchmark for this approach.

Following the development of Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001), Newman et al (2003) attempted to automate deception detection by using the psycho-linguistic word categories to create a linguistic profile of deception, demonstrating the promise of such an approach. Computational techniques include Zhou et al (2004) , who studied the automation of linguistic based cues on text-based computer mediated communication. The work of Hancock et al (2008) further explored computer mediated communication by analyzing the words of the partner as well as the deceiver, to achieve an improved classification rate.

In 2011, Ott et al. (2011) developed a gold standard deceptive data-set of truthful and deceptive hotel reviews for their work on Deceptive Opinion Spam, and achieved promising results in the task of Spam Detection. Ott et al. (2013) further expanded upon this data-set to include negative opinion spam.

Recently, multimodal approaches have been introduced to improve the accuracy of detecting deceit. Features like visual, acoustic, text and micro expressions are integrated to analyze the result. Data used in these approaches is either collected through lab settings or real life trial data, collected and made available by Pérez Rosas et al (2015).

Another study by Abouelenien et al (2017) integrated multiple physiological, linguistic, and thermal features. The results showed that a multimodal approach can be a promising step towards an automated system to detect deceit. Mendels et al. (2017) aimed at automatically detecting deception from speech and conducted a series of experiments. They compared the use of acoustic-prosodic and lexical feature sets, using various machine learning models and designed a single hybrid deep model with both acoustic and lexical features trained jointly.

Krishnamurthy et al. (2018) in their approach analyzed multimodal video data, by first extracting unimodal features from each video. The features used in this approach were textual, audio and visual. To integrate the features, data fusion techniques: concatenation and Hadamard product with concatenation were used.

The results of the multimodal approaches are indicative of the significance of visual and textual features, primarily, to the accurate prediction of deception. In this paper, we aim to specifically analyze the significance of textual features in a deception model.

III. Datasets

One of the major issues plaguing deception studies is the lack of readily available labelled data. In our study, the following datasets are used:

A. Deceptive Opinion Spam Corpus

Ott et al. (2011) developed a deceptive opinion spam dataset, with gold-standard deceptive opinions, by collecting truthful and deceptive positive online hotel reviews on the 20 most popular Chicago hotels according to TripAdvisor. All reviews are at least 150 characters long.

• The **deceptive** reviews were produced through crowd-sourcing using Amazon Mechanical Turk, where 400 Turkers were paid

to submit believable fake online reviews on the relevant hotel (20 on each hotel), within a time limit of 30 minutes.

• The positive (5-star) **truthful** reviews were mined from TripAdvisor on the 20 relevant hotels, resulting in more than 2000 reviews. Of these, 20 were selected on each hotel from a log-normal distribution fit to the lengths of the deceptive reviews.

Thus, the dataset consisted of 400 deceptive and 400 truthful positive hotel reviews.

Ott et al. (2013) then expanded upon this to include negative opinion reviews using the same technique as the positive review collection procedure. Thus, they obtained 400 deceptive negative reviews from Mechanical Turk, and 400 truthful negative reviews from six popular online review communities (Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp).

The final combined corpus thus contains 800 positive and 800 negative reviews.[3]

B. Real Life Trial Dataset

Pérez-Rosas et al. (2015) (2015) created a dataset consisting of real-life trial videos that are publicly available on YouTube channels and other public websites. The dataset also contains statements made by exonerees after exoneration and a few statements from defendants during crime-related TV episodes. The speakers in the videos are either defendants or witnesses. The video clips are labeled as deceptive or truthful based on a guilty verdict, not-guilty verdict, and exoneration.

Figure 1: A sample still of a deceptive trial video

The final dataset consists of 121 trial videos, of which 61 are deceptive and 60 are truthful. There are 21 unique female and 35 unique male speakers. The average video length is 28 seconds.

The **transcripts** for the videos were generated through Amazon Mechanical Turk, and manually verified to ensure their quality. The final set has an average word count of 66 per transcript.

IV. Methodology

1.Traditional Machine Learning Classifiers - The usage of textual features have shown promise in past works (Mihalcea & Strapparava, 2009; Ott et al., 2011) when used to train classifiers such as Support Vector Machine and Naive Bayes Classifiers (Zhou et al., 2008) for automated deception detection. We attempt a similar approach.

2. Deep Learning Model - As demonstrated by Mendels et al. (2017), who used a bidirectional Long Short Term Memory (LSTM) model, as well as Krishnamurthy et al. (2018) who used a Convolutional Neural Network (CNN), deep learning models show promise in the task of deception detection in text. We hence utilize similar models for our task and evaluate their performance.

The spam detection dataset (Dataset 1) consists of written (typed) data whereas the Real-Life data (Dataset 2) is essentially transcripts of spoken language. We apply our traditional classifier approach (Approach 1) across both datasets and gauge their comparative performance. While we expect that, due to the lack of instances (121) in the Real-Life trial dataset, we will be able to achieve a better performance on the Spam Dataset, we seek to analyze whether our techniques are able to perform across both types of data. While we aim to gauge how our approach performs on Dataset 1 as well, in particular, we wish to evaluate how our classifiers perform on Dataset 2, as we do have an indication from past studies that these classifiers are able to per-

[3] The original corpus can be found at http://myleott.com/op_spam/

form well on dataset 1 (Ott et al., 2011; Ott et al., 2013).

We evaluate the performance of our deep learning models (Approach 2) for the task of spam detection (on dataset 1), as Dataset 2 is deemed to be too small for accurate evaluation of a deep learning model, and prone to overfitting.

We shall delve further into the different techniques in this section, beginning with the textual features used in Approach 1, as follows:

A. Text Features

1.LIWC

The Linguistic Inquiry and Word Count (LIWC) software is a popular automated text analysis tool used broadly in the social sciences. It has been used to detect personality traits, to study tutoring dynamics, and, most relevantly, to analyze deception (Hancock et al., 2008). While LIWC does not include a text classifier, we can create one with features derived from the LIWC output. In particular, LIWC counts and groups the number of instances of nearly 6,400 words, word stems, and emoticons into around 90 psychologically meaningful dimensions. We construct one feature for each LIWC dimension. We have made use of the most recent, *LIWC2015* dictionary to produce the LIWC features. The classification results in Table 1 are based on 93 LIWC dimensions from the *LIWC2015* dictionary, which contains more categories, and more words than the earlier *LIWC2001*, or *LIWC2007* versions.

As found by Mihalcea et al. (2009), the 5 "most-significant categories" for Deceptive class (METAPH, YOU, OTHER, HUMANS, CERTAIN) and Truthful class (OPTIM, I, FRIENDS, SELF, INSIGHT), from the *LIWC2001* dictionary were used for classification. Focusing on these categories, however, did not give an improved accuracy for classification, as compared to when the classifier was allowed to train over all the LIWC features freely.

As found by Mihalcea et al. (2009), the 5 "most-significant categories" for Deceptive class (METAPH, YOU, OTHER, HUMANS, CERTAIN) and Truthful class (OPTIM, I, FRIENDS, SELF, INSIGHT), from the *LIWC2001* dictionary were used for classification. Focusing on these categories, however, did not give an improved accuracy for classification, as compared to when the classifier was allowed to train over all the LIWC features freely.

2. Text Categorization

A text categorization approach to deception detection includes both content and context with n-gram features. We consider the following three n-gram feature sets: UNIGRAMS, BIGRAMS*, TRIGRAMS*, where the * indicates that the feature set contains the previous feature set. We analyze the *tf-idf* (term frequency-inverse document frequency) features of the text documents for each of the NGRAM feature sets. *tf-idf* is a statistic that indicates the importance of a term or phrase to a document within a corpus. The goal of using tf-idf instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus. tf-idf is computed as the product of two measures:

- Term Frequency - As the name suggests, it refers to the frequency of a particular term in a document.

- Inverse Document Frequency - It is a measure of how common or rare the term is across all documents. The idea is that a word that occurs commonly across all documents, such as 'the', may not be very relevant, yet a term that is unique to a particular document is highly useful in distinguishing the document from others. As we use simple smoothing to deal with zero frequency terms, the formula for idf becomes:

$$idf(d,t) = log[\frac{(1+n)}{(1+df(d,t))}] + 1$$

- In the above formula n is the total number of documents in the document set and df(d,t) gives us the frequency of the term(t) in each document(d) in the corpus.

As preprocessing of the text data, we eliminate stopwords that are extremely common across all documents, such as 'a', 'an', 'the' etc. How-

ever, we do not perform stemming and lemmatization of words as we found these techniques to reduce the performance of classification on both the datasets.

B. Traditional Classifiers

1. As demonstrated in existing work (Mihalcea et al., 2009; Zhou et al., 2008), Naive-Bayes and Support Vector Machine Classifiers are shown to perform well when trained on similar features as mentioned above[4]. We further examine the performance of a Random Forest classifier on our feature sets. Thus, we tabulate results for the following:

- Naive Bayes Classifier (NB) (Friedman et al., 1997)
- Support Vector Machine Classifier (SVM) (Cortes & Vapnik, 1995)
- Random Forest Classifier (RF) (Breiman, 2001)

We train our classifiers on the different feature sets, namely UNIGRAMS, BIGRAMS*, TRIGRAMS* and LIWC features, as well as on combinations of the above. We make use of a train-test split of 90%-10% and compute the evaluation metrics, which are listed in Table 1 and 2.

C. Deep Learning Models

Deep Learning is a subset of Machine Learning, which deals with algorithms based on the structure and operation of the human brain, called *artificial neural networks*.[5] The significant advantage of deep learning, over traditional machine learning, is that the problem must be broken down into a simpler form for a machine learning algorithm, whereas a deep learning algorithm is able to solve this problem itself. This means that for a machine learning algorithm, we must first perform relevant feature extraction on our data, and then feed these features into the ML classifier (as we have seen in the previous section). A deep learning network aims to perform its own feature extraction and classification. Hence, it solves the problem "end to end".

1. Convolutional Neural Network (CNN)

Convolutional Neural Networks are a special category of neural networks designed primarily for image processing (see Appendix C-A). However, more recently, they have been applied directly to text analytics. As demonstrated by Kim (2014), a simple CNN can perform extremely well for various NLP tasks, such as sentiment analysis and text classification. We hence use a similar CNN model for our task of deceptive review detection. The model first extracts features from our text data (through a process called convolution), and then uses them to classify the instances as truthful or deceptive. As input to the CNN, we need to extract the vector representations of the individual words in our text data, known as word embeddings [Appendix C-A1]. We have done this in three different ways:

- Embedding Layer, initially randomized and trained along with the CNN
- Google's pre-trained word2vec model (Mikolov et al, 2013).
- Google's pre-trained BERT[6]model (Devlin et al, 2019).

We have documented the performance of our model with each of the different embedding techniques in Table 4.

2.LSTM Network

Long Short Term Memory Networks (Hochreiter & Schmidhuber, 1997) are a special kind of artificial neural network, that are able to remember "long-term dependencies" in data, i.e. they are able to draw context from relatively earlier data, such as a significant word from a previous sentence. E.g. In the sentence, "I

[4]See Appendix B for additional information on classifiers.

[5]Refer Appendix C for more information on neural networks.

[6] Bidirectional Encoder Representations from Transformers - BERT is the "first *deeply bidirectional, unsupervised* language representation, pre-trained using only a plain text corpus". It has achieved state-of-the-art results in 11 Natural Language Processing Problems.

grew up in France. I speak fluent ..." (Olah, 2015), an LSTM will be able to predict the next word as *French* by drawing context from the previous sentence. Thus, LSTMs are able to overcome the drawback of traditional Recurrent Neural Networks, which are unable to learn long-term dependencies. (See Appendix C-B).

We use 85% of the data to train the model and 15% to test it.

V. Results

A. Machine Learning Methods

1.Deceptive Opinion Spam Corpus

Table 1 shows the performance of various classifiers and features on the Deceptive Opinion Spam Corpus. The results show that models trained only on UNIGRAMS - the simplest n-gram feature set, are better than the psycho-linguistic deception detection approach. Further the results with BIGRAM features are even better. This suggests that an individual keyword-categorization based approach to deception detection (eg. LIWC) is not the most optimal (best performance using LIWC = *79%*)[7], and an approach focusing on the context (eg. BIGRAMS) might be necessary to achieve a better performance.

We have used Support Vector Machine, Naive Bayes and Random Forest classifiers in our approach along with feature sets which include UNIGRAM, BIGRAM, TRIGRAM and LIWC. We have analyzed every combination and the accuracy and F1 score obtained by each combination is stated in Table 1. We see that SVMs tend to achieve the best performance while Random Forest Classifiers do not appear to work well on our task.

For combining LIWC and NGRAMs, we generate the feature vectors for LIWC and NGRAMS individually, then unit length normalize the two vectors and concatenate them for classification. The accuracy and F1-Score for this approach can be seen in Table 1. We perform an optimization of the number of features supplied to the classifier in order to maximize performance and arrive at a value of 1600 BIGRAM features in addition to the 93 LIWC features, for this particular combination. While using only BIGRAM features, we require 3200 features to achieve the best accuracy.

Ott et al. (2011) showed that an inclusion of LIWC features along with BIGRAM features was the best indicator of deception achieving an accuracy of 89.8%, a marginal improvement from the 89.6% they achieved with only BIGRAM features. However, we are able to achieve an accuracy of 91.88% using only BIGRAM features, and experience a slight decrease in accuracy when combining them with LIWC (91.25%), although this combination is only marginally second best.

While the study of Ott et al. (2011) used only positive hotel review data, a follow up study (Ott et al., 2013) attempted to analyze both positive and negative-modality reviews, achieving a best accuracy of 88.4% when training over the entire combined dataset, as we have. Thus, the usage of smoothed tf-idf BIGRAM features demonstrate an improved performance in the categorization of deceptive spam.

[7] See Appendix A for more information on statistical measures used

Table 1: Automated classifier performance for two approaches on Online Spam Dataset

Approach	Features	Accuracy	F1 - Score
PSYCHO LINGUISTIC DECEPTION DETECTION	LIWC(SVM)	79.1%	79%
	LIWC(NB)	65.4%	70%
	LIWC(RF)	67.9%	59.6%
TEXT CATEGOR-IZATION	BIGRAM(SVM)	91.88%	90.91%
	BIGRAM(NB)	85.6%	83.6%
	BIGRAM(RF)	68.7%	62.1%
	LIWC+UNIGRAM (SVM)	88.75%	87.67%
	LIWC+BIGRAM (SVM)	91.25%	90.28%
	LIWC+TRIGRAM (SVM)	90.00%	89.04%

Table 2: Automated classifier performance for two approaches on Real Life Trial Dataset

Approach	Features	Accuracy	F1-Score
PSYCHO LINGUISTIC DECEPTION DETECTION	LIWC(SVM)	57.8%	60%
	LIWC(NB)	63.1%	70.99%
	LIWC(RF)	52.6%	40%
TEXT CATEGORIZATION	UNIGRAM(SVM)	73.7%	78.3%
	UNIGRAM(NB)	68.4%	66.7%
	BIGRAM(SVM)	76.9%	80%
	BIGRAM(NB)	69.2%	71.4%
	BIGRAM(RF)	46.1%	53.3%
	LIWC + BIGRAM(SVM)	76.9%	80%

2. Real Life Trial Dataset

Table 2 shows the performance of classifiers like SVM, Naive Bayes, Random Forest with psycho-linguistic and text categorization features on the Real Life Trial Dataset. This dataset has not previously been analyzed with the approaches stated in Table 2. The results in this part also clearly show that the accuracy increases when the features focus more on the context of the dataset than the keywords. The accuracy obtained is the highest when Support Vector Machine classifier is used with BIGRAM features.

We have analyzed different combinations and the accuracy obtained by each is stated in Table 2. In line with our results on the spam dataset, we find the SVM classifier to give the best results while the Random Forest classifier is unable to predict with an accuracy significantly better than chance. For combining LIWC and NGRAM features the same steps explained in the Deceptive Opinion Spam Corpus are repeated. The optimal number of BIGRAM features for this combination are found to be 2000, as compared to 1600 on the spam dataset.

The highest accuracy of 76.9% is obtained by an SVM classifier trained on BIGRAM tf-idf features (3000 features), and a similar accuracy is obtained by including the combination of LIWC and BIGRAM features for Support Vector Machines. In accordance with the results on the Spam Dataset above, we find that inclusion of the psycho-linguistic features does not contribute to an increase in accuracy. However, the LIWC + BIGRAMs combination is able to achieve the same accuracy with fewer features as compared to just BIGRAM features (2093 as compared to 3000).

The difference in the corresponding accuracy values over the 2 datasets is likely due to the difference in the size of the datasets, as was postulated. The Real-Life Trial dataset contains far fewer instances (121) in comparison to the Spam Detection dataset (1600), and thus does not give as high an accuracy. With an increase in quality data, it is likely that prediction of deception from text will deliver a better performance. However, we can clearly see that the text feature analysis approach is able to distinguish between deceptive and truthful transcripts of spoken language with an accuracy significantly greater than chance.

Table 3: Number of BIGRAM features for optimal accuracy with SVM

Dataset	Approach	Total number of Features	Accuracy
Spam Detection Data	LIWC + BIGRAMS	1693	91.25%
	BIGRAMS	3200	91.88%
Real Life Trial Transcripts	LIWC + BIGRAMS	2093	76.9%
	BIGRAMS	3000	76.9%

B. Deep Learning Approaches

As mentioned earlier, we implement our deep learning models for the task of spam detection, and not on the trial transcripts dataset. The results of our two models are listed in Table 4. It is clear that we are able to achieve a significant accuracy in correctly classifying the spam data with each of our deep learning approaches. With our CNN model, we achieve a best accuracy of 87.5% when training the word embeddings ourselves through an embedding layer. The inclusion of pretrained embeddings from Google's word2vec model gives us a marginally better accuracy, and the usage of BERT embeddings improves this even further. Our best accuracy of 90% is comparable to our best result in Table 1, although we do not provide our CNN model with any prior extracted features (such as BIGRAM tf-idf counts, LIWC etc, which we do provide in our ML classifier-based approach). Our LSTM model achieves an accuracy of 84.6% on the spam dataset. We find that the inclusion of pretrained embeddings does not improve the accuracy of our LSTM model. Thus, our Deep Learning approaches are successfully able to learn the task of classifying online spam, with an accuracy comparable to our best Machine Learning Approaches. Generally, deep learning models are able to scale better given more data, and outperform their machine learning counterparts after a certain threshold i.e. the more the data, the better performance for a deep learning model. We believe that with a larger dataset, our deep learning models will perform even better in the task of online deception detection. Further, the fact that Machine Learning models often require complex feature extraction and engineering, highlights the convenience and promise of deep learning, in this domain.

Table 4: Deep Learning Classifiers - Performance on Spam Dataset

Model	Embedding Type	Accuracy
CNN	Embedding Layer	87.5%
	Word2vec embeddings	88.75%
	BERT embeddings	90.0%
LSTM	Embedding Layer	84.6%

VI. Future Direction

A. Exploring Adversarial AI

1. Background

Neural networks, despite their significant task performance and inspiration from biological systems, often behave in ways that challenge the intuition of human observers; signaling a peculiar fragility and brittleness present in the models that are anticipated to take over the field. This brittleness is often highlighted in the context of model sensitivity to very specific changes in inputs known as adversarial examples. In simpler terms, adversarial examples are inputs that are specifically crafted to fool machine learning classifiers. In most cases, adversarial examples are indistinguishable from their "real" counterparts to humans. The nature of adversarial examples is quite concerning since an attacker could feed inputs that are designed to yield a certain predefined outcome from the machine learning model.

It turns out that adversarial examples generated from a specific model are transferable to other machine learning classifiers as well. This property, referred to as "transferability of adversarial examples" is particularly problematic because malicious inputs could cause significant damage when high stake decisions are made or driven by machine learning models. Fortunately, attacking text based models is more complicated than attacking image classification models due to the form of textual input. Text based models map each word in the input to a vector, referred to as Word Embeddings. Due to the discrete mapping of each word to a fixed vector, it is hard to craft adversarial inputs without making them different from the source inputs. So, adversarial inputs in the text domain focus instead on retaining the semantic information in the input. We craft adversarial examples by replacing a chosen set of words with the words that correspond to the nearest embeddings.

It is important to note that adversarial attacks are a vulnerability for all neural network models, we do a brief exploration here to see whether a basic adversarial attack is able to thwart our classifiers. We plan to explore this further in the future.

2. Initial Findings:

We use the CNN mentioned in Section IV-C1 to craft adversarial examples. By following the process described above, we were provided with interesting insights into the dynamics and learnings of the deep learning model. We demonstrate the results of the word replacement by crafting the adversarial example from a truthful review

Original sample: *my husband and i arrived at the swissotel chicago to celebrate our 13th wedding anniversary. Upon arrival at the given and advertised check in time* ***our*** *room was not ready we waited for more than an* ***hour with*** *our bags.*

Sample with the nearest embedding replacement: *my husband and i arrived at the swissotel chicago to celebrate our 13th wedding* ***it****. Upon arrival at the given and advertised check in time* ***they*** *room was not ready we waited for more than an* ***were up*** *our bags.*

The words highlighted correspond to the words that have been replaced with the nearest embed- dings. It is interesting to note that the CNN embeddings layer has found the word it to be the one closest to the word anniversary. It is clear that the crafted example is not semantically consistent, yet, the model classifies this as a truthful review with a very high confidence of 0.99, the same result as the original sample. Hence, our model is not fooled despite the input being tampered with.

VI. Conclusion

In this study we aimed to explore the problem of human deception detection: in particular, how the language used by a human while attempting to deceive can be analyzed to illustrate deception. We used two different sets of data; one consisting of online hotel reviews, both real and fake, and one consisting of transcripts of clips of real court trials, each marked as either truthful or deceptive. We wished to evaluate, one, whether our textual analysis approach is relevant to both online text as well as transcribed spoken text, and two, whether a deep learning model is suited for this task. For the former, we analyzed text features such as n-gram tf-idf, and LIWC features, and trained classifiers on these features to predict deceptiveness in text. We attempted different combinations of the features along with different classifiers in order to achieve the best accuracy of prediction.

We found that a Support Vector Machine Classifier was able to achieve the highest accuracy, followed by a Naive Bayes Classifier. The Random Forest classifier was found to perform poorly for this task. We also found the optimal number of features for each feature set and dataset combination, and have listed the most significant ones in Table 3. For the feature sets, the best accuracy of 91.88% on the spam dataset, and of 76.9% on the Real-Life Trial dataset, are both achieved with BIGRAM tf-idf features trained using an SVM classifier. Thus we see that BIGRAMs are able to outperform the TRIGRAM feature set. We also find that, while the inclusion of LIWC features along with n-gram features does not necessarily deliver a better accuracy, it is able to achieve almost as good (on spam data), if not equal accuracy (on trial transcripts), with a significantly fewer number of features (refer Table 3).

As a deep learning approach, we developed a Convolutional Neural Network and an LSTM network for the task of spam detection. Our CNN model, with an accuracy of 90%, outperformed our LSTM model (84.6%), although we can deem both successful in the task of spam detection. We highlight the convenience of deep learning models, and the likelihood of improved performance with a larger dataset.

We are successfully able to achieve a considerable accuracy in the task of spam detection from hotel reviews, an improvement from the previous results on this dataset , and also on the Real-Life Trial data transcripts, with an accuracy significantly better than chance. Our results demonstrate that the usage of textual features shows significant promise in the task of deception detection from language, whether it be online spam or transcriptions of spoken language. We also show that deep learning methods - in particular CNN based models - are promising in the task of online fraud detection from text.

VII. Acknowledgments

The authors would like to thank V. Cholan and Don Krapohl for their reviews and edits of earlier manuscripts. The authors would also like to thank Temasek Laboratories @ NTU, Singapore, without whom this project would not have been possible.

References

Abouelenien, M., Pérez-Rosas, V., Burzo, M., & Mihalcea, R. (2017). Detecting Deceptive Behavior via Integration of Discriminative Features from Multiple Modalities. *IEEE Transactions on Information Forensics and Security, 12*(5), 1042-1055. https://doi.org/10.1109/TIFS.2016.2639344

Breiman, L. (2001). Random Forests. *Machine Learning,* 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Britz, D. (2015, November 7). *Understanding Convolutional Neural Networks for NLP.* WildML. http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. *Intelligence and Security Informatics,* 2665, 91–101. https://doi.org/10.1007/3-540-44853-5_7

Cortes, C., & Vapnik, V. (1995). Support-vector Networks. Machine Learning, 20, 273-297. https://doi.org/10.1007/BF00994018

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129(1), 74-118. https://doi.org/10.1037/0033-2909.129.1.74

Devlin, J., & Chang, M-W. (2018, November 2). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Google AI Blog. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Donges, N. (2018). *The Random Forest Algorithm.* Towards Data Science. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning,* 29, 131-163. https://doi.org/10.1023/A:1007465528199

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Processes, 45(1), 1-23. https://doi.org/10.1080/01638530701739181.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(*8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1746–1751. https://doi.org/10.3115/v1/D14-1181

Krishnamurthy, G., Majumder, N., Poria, S., & Cambria, E. (2018). A Deep Learning Approach for Multimodal Deception Detection. *arXiv preprint.* arXiv:1803.00344

Mendels, G., Levitan, S. I., Lee, K-Z., & Hirschberg, J. (2017). Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. *Proc. Interspeech 2017,* 1472–1476

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers,* 309-312.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Jeffrey, D. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *In Proc. Advances in Neural Information Processing Systems,* 2, 3111–3119.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin,* 29(5), 665–675. https://doi.org/10.1177/0146167203029005010

Nielsen, M. A. (2015). *Neural Networks and Deep Learning – Chapter 1.* Determination Press.

Olah, C. (2015, August 27). *Understanding LSTM Networks.* https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. *Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 497–501.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1*, 309– 319.

Patel, S. (2017, May 11). *Decision Tree Classifier* — Theory. Medium. https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

Patel, S. (2017, May 18). *Random Forest Classifier.* Medium. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1

Patel, S. (2017, May 3). SVM *(Support Vector Machine)* — Theory. Medium. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *17th ACM International Conference on Multimodal Interaction,* 59-66. https://doi.org/10.1145/2818346.2820758

Pupale, R. (2018, June 16). *Support Vector Machines* (SVM) — An Overview. Towards Data Science. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Saha, S. (2018, December 15). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way.* Towards Data Science. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Soni, D. (2018, May 17). *Introduction to Naive Bayes Classification.* Towards Data Science. https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting Deceit via Analysis of Verbal And Nonverbal Behavior. *Journal of Nonverbal Behavior, 24*(4), 239-263. https://doi.org/10.1023/A:1006610329284

Vrij, A., Leal, S., & Fisher, R. P. (2018). Verbal Deception and the Model Statement as a Lie Detection Tool. Front. *Psychiatry*, 9, 492. https://doi.org/10.3389/fpsyt.2018.00492

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-based Asynchronous Computer-Mediated Communication. *Group Decision and Negotiation, 13,* 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

Zhou, L., Shi, Y., & Zhang, D. (2008). A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Transactions on Knowledge and Data Engineering,* 20(8), 1077–1081. https://doi.org/10.1109/TKDE.2007.190624

APPENDIX A

STATISTICAL MEASURES USED

In order to evaluate the performance of our models, we use 2 main statistical measures: Accuracy and F1 score.

To better understand these metrics in terms of classification, we must introduce 4 count measures:

- True Positives (TP) - This is the number of samples which are correctly classified as positive for a class, i.e. the true class value of the samples is positive and the predicted class is also positive.

- False Positives (FP) - This is the number of samples which are incorrectly classified as positive for a class, i.e. the true class value of the samples is negative but the predicted class is positive.

- True Negatives (TN) - This is the number of samples which are correctly classified as negative for a class, i.e. the true class value of the samples is negative and the predicted class is also negative.

- False Negatives (FN) - This is the number of samples which are incorrectly classified as negative for a class, i.e. the true class value of the samples is positive but the predicted class is negative.

A. Accuracy

The accuracy of a class is a simple performance metric, which can be defined as the ratio of the number of correctly predicted samples to the total number of samples.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

The drawback of accuracy as a metric is that it does not capture the true performance of a classifier when given an uneven class distribution. For example, consider a particular distribution of 100 samples, consisting of 90 samples from class A and 10 from class B. A classification model may be biased towards class A and simply predict class A for every single sample. This model will be able to achieve an accuracy score of 90%, which indicates a very good performance. However, we know that this classifier would not work well for a different data distribution. Thus, the accuracy measure is not able to effectively encapsulate the true performance of a classifier in such a case. F1-score is able to overcome this drawback and better capture the overall performance of the classifier.

B. F1-score

F1-score is a performance metric that, like, accuracy varies between 0 and 1 (1 being a perfect classifier). It is defined in terms of two metrics:

- Precision – Precision, or Positive Predictive Value (PPV), refers to how 'precise' a model can be termed as, i.e. out of the values predicted as positive by the model, how many of them are actually positive? Thus it can be defined as:

$$Precision = TP / (TP + FP)$$

- Recall - Recall refers to the ability of a model to recall correctly all of the positive samples i.e. out of all the values that are actually positive, how many of them are recalled by our model and correctly classified as positive? It is defined as:

$$Recall = TP / (TP + FN)$$

F1-score seeks to combine precision and recall to give a comprehensive performance metric. It is defined as the harmonic mean of precision and recall, and hence can be mathematically written as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In our study, we have looked at both accuracy, as well as F1-score, to evaluate our classifiers.

While Negative Predicted Value (NPV) is a statistic often used in credibility assessment, it is not used while evaluating our automated classifiers in this study – a combination of PPV and Recall (namely F1 score) enables us to understand best how our classifier is performing. NPV can be seen as the Precision with which the classifier predicts the negative values, and it could be interesting to incorporate this in future work.

APPENDIX B

CLASSIFIERS

In Machine Learning, classification is the task of predicting the 'class' of a data point using the information gained from previous data points. For example, one may analyze the outside temperature for several days as being termed cold, pleasant, or hot, and note the corresponding temperature value. Here, there are 3 classes - cold, pleasant, and hot, that our data points (or temperature readings) fall into. Now, on a new day, when one takes a temperature reading, one can judge which class it should fall into based on our earlier knowledge of similar temperatures. Thus, one can classify the temperature data into one of the three classes. In our study, we attempt to perform a two-class classification where the two classes are truthful and deceptive. Based on the textual or psycholinguistic features of our input, we attempt to decide whether it belongs to the truthful or deceptive category.

A model that can perform the task of classification is broadly known as a classifier. There are several different types of classifiers, each employing different techniques to achieve the same goal of classification. In our study, we make use of three different classifiers, and evaluate their performance on our task.

A. Naive-Bayes Classifier

The Naive-Bayes classifier is an uncomplicated, yet efficient and popular classifier, built upon the simple but powerful Bayes' Theorem, that can be written as follows:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

where A and B are events, and. $P(B) \neq 0$. P(A) represents the probability of event A occurring, P(B) is the probability of event B occurring, and P(A|B) is the conditional probability of event A occurring given that B has occurred. This rule is used to model the class probability, given a set of features as follows:

$$P(C_i|x_0, x_1, \ldots, x_{n-1}) = \{P(x_0, x_1, \ldots, x_{n-1}|C_i) * P(C_i)\}/\{P(x_0, x_1, \ldots, x_{n-1})\}$$

Here, C_i represents a particular class, while x_0,x_1, etc represent each of the n features of the input data. This rule is rewritten with two changes for convenience:

- Naive approximation - We make the assumption that, given C_i, the features x_0,x_1, etc are conditionally independent. Hence, we can write. $P(x_0, x_1, \ldots, x_n - 1|C_i) = P(x_0|C_i) \times P(x_1|C_i) \times \ldots \times P(x_{n-1}|C_i)$ This is known as the 'naive' approximation. Despite this inexact assumption, the model is known to work well as a classifier.

• The denominator is not of significance and we eliminate it

Thus, we obtain the expression:

$$P(C_i|x_0, x_1, \ldots, x_{n-1}) \propto P(C_i) \times \prod_{j=0}^{n-1} P(x_j|C_i)$$

In order to classify a data point, we simply choose the class with maximum probability, given the set of features of the data point. i.e. we pick C_{max} such that:

$$C_{max} = arg\,max_{C_i} P(C_i|x_0, x_1, \ldots, x_{n-1})$$

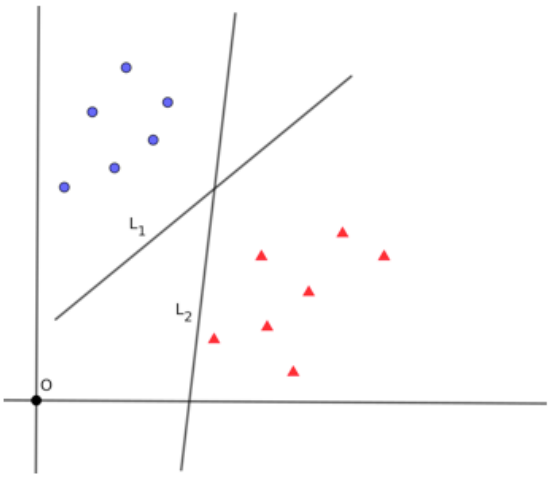
This rule is known as the Maximum A Posteriori decision rule, and is essentially exactly what the classifier uses to predict the class for a particular data instance.

Variants of the Naive Bayes classifier differ by the assumptions they make about the distribution of $P(x_j | C_i)$ A Gaussian Naive Bayes classifier assumes each feature to be distributed as per the Gaussian distribution (also known as a normal distribution). Some other types include the Multinomial Naive Bayes and the Bernoulli Naive Bayes classifiers. In our study, we use the Multinomial Naive Bayes classifier as it is found to be the most suitable.

B. Support Vector Machines

A Support Vector Machine (SVM) is a classification model that performs well for not only linear but also non-linear problems. The concept is straightforward: the algorithm constructs a line, or a *hyperplane*[8], in order to divide the data into classes. Let us consider a simple, two class problem as seen in Figure 2.

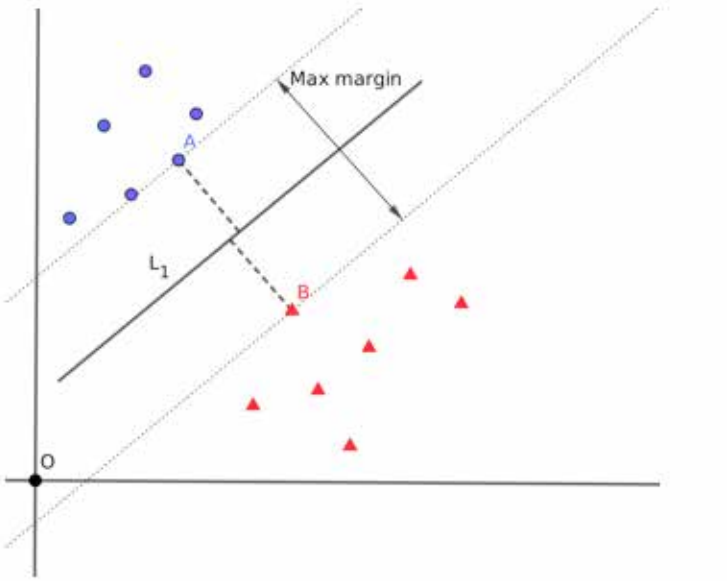
Figure 2: Two lines of separation for the two classes



[8] As per (Patel, 2017): "A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts." In the demonstrated examples, our hyperplane is a single-dimensional line (whether straight or circular), that is able to separate our 2-dimensional data distribution into distinct classes.

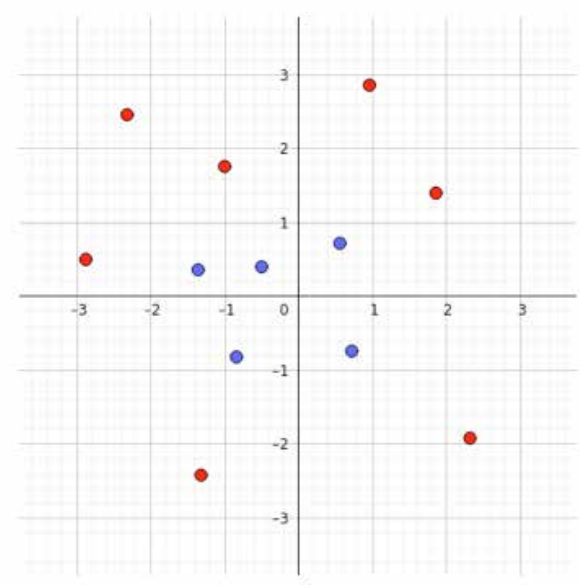
The SVM algorithm attempts to separate the two sets of data points, or classes (blue v/s red), with a single line. However, it is clear that there are several different lines that would be able to do the needful. In order to choose the best line of separation, the algorithm looks for the data points closest to the line (we call them 'support vectors') and calculates their perpendicular distance (called the margin) from the line of separation. In Figure 3, points A and B are support vectors. The best line of separation is the one which maximizes the margin (L_1 this case).

Figure 3: L_1 is the line of best fit



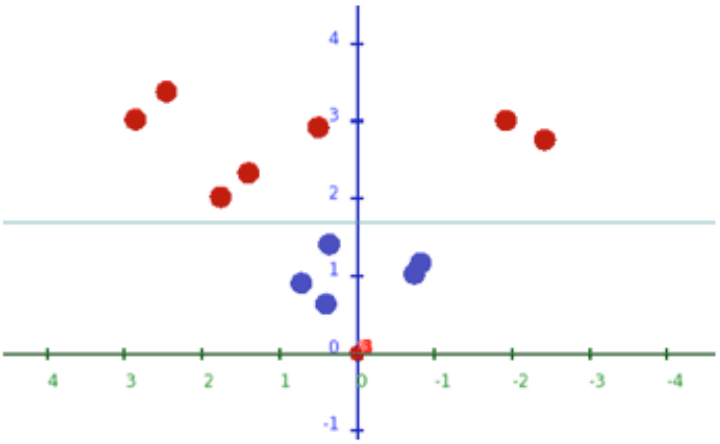
Classes are not always linearly separable. Take Figure 4 for instance. It is evident that a straight line would be unable to separate the 2 classes. However, this data can be transformed such that it is linearly separable in a higher dimension.

Figure 4: The two classes are not linearly separable



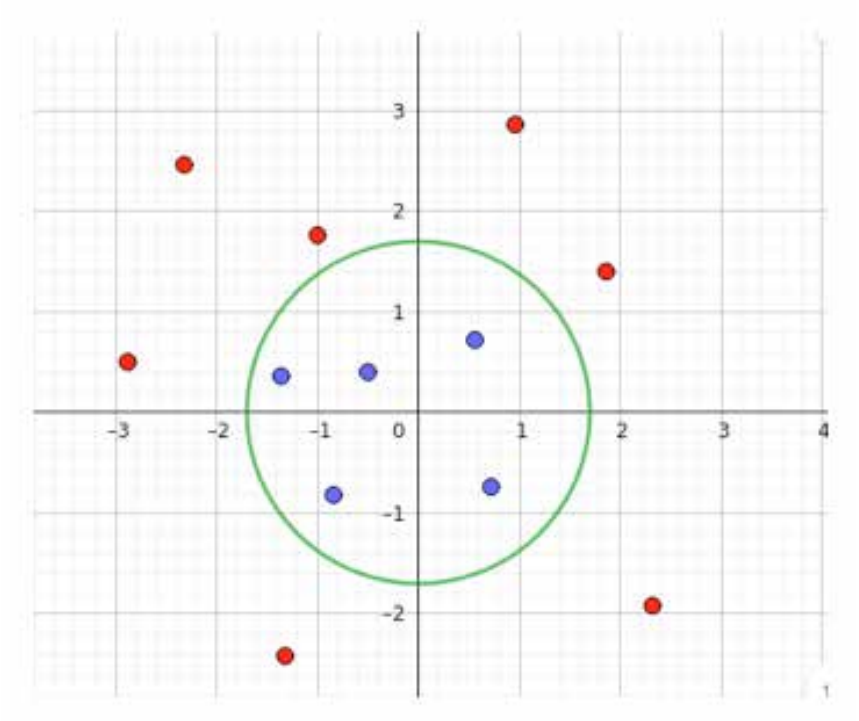
Let us introduce the Z-dimension. We may define the z-coordinate for each data point as the square of the distance from the origin i.e. $z = x^2 + y^2$, for each point (x,y). This would transform our distribution to Figure 5 in the Z-dimension.

Figure 5: Transforming distribution to the Z-dimension



We see that this is now linearly separable. The decision boundary, or line of separation, can be given by the equation $z = k$ .When we transform this back to our original 2 dimensions, we arrive at the equation $x^2 + y^2 = k$. This is the equation of a circle, and our decision boundary can be seen in Figure 6.

Figure 6: Decision Boundary



Thus, an SVM classifies data by transforming it into a higher dimension such that it becomes linearly separable, and then maps the higher dimension line back to the original dimension, establishing an accurate decision boundary. These transformations are performed with the help of a set of mathematical functions, known as the 'kernel'. There are differing kernel functions used by different SVMs, such as linear, non-linear, polynomial, sigmoid etc. We have utilized a linear kernel in our study.

C. Random Forest Classifier

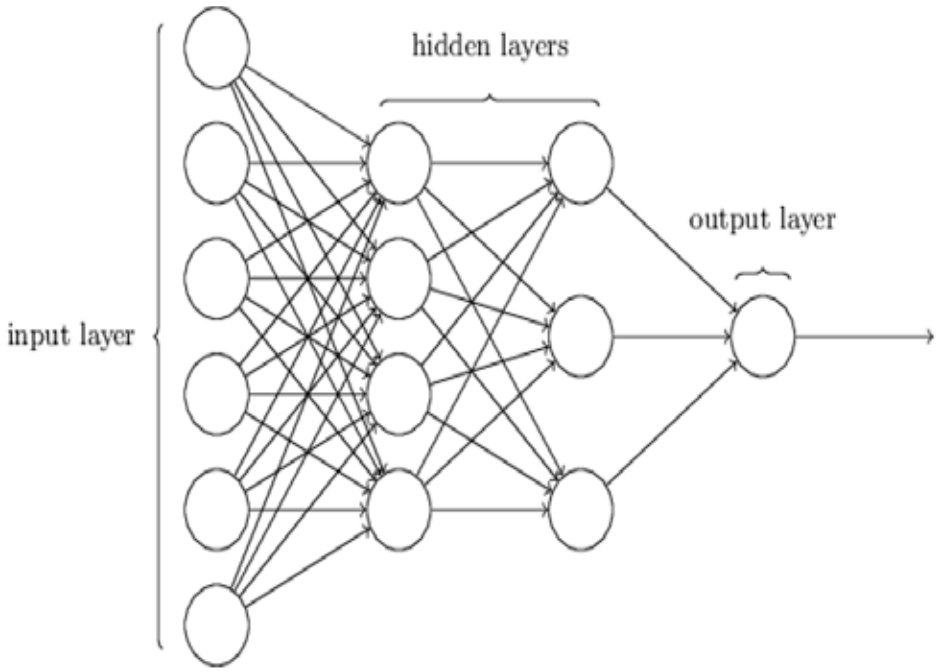
The Random Forest classifier can be termed as an "ensemble algorithm", i.e. it incorporates features from one or more algorithms. A Random Forest is comprised of a collection of Decision Trees[9], and makes the final classification decision based on a combination of votes from the individual decision trees. The final aggregate decision can be made either by picking the majority, or by assigning weights to the votes of the individual trees (E.g. A lesser weight to a tree with a high error rate) and computing the net result. Random Forest classifiers vary from each other in terms of the number of trees, and the individual decision tree parameters like maximum depth, minimum samples required to split a node, etc. A great advantage of a Random Forest classifier is that it prevents over-fitting (learning only a particular dataset), due to the large number of different decision trees. However, while a larger number of trees would result in a more accurate classification, it can make the algorithm too slow for real-time computation.

APPENDIX C

ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a computing system built to loosely replicate the neural network structure of the human brain. It consists of a collection of nodes, or neurons (based on biological neurons), that can transmit signals between each other through connections, which are based on the synapses in the brain. These nodes are organized in various layers, as can be seen in Figure 7.

Figure 7: A Typical Neural Network Source



[9] A Decision Tree Classifier is an algorithm that generates a set of rules or decisions, organized in a hierarchical tree like fashion. The final output class labels are at the leaves (the final level) of the tree. While attempting to classify a data instance, a 'decision' is made at each rule node of the tree, as to which of its child branches to explore, until the final level, and hence class output, is reached.

On receiving an input signal, each layer of neurons performs its computation and passes a signal (or a numerical value) on to the neurons of the next layer. This continues until an output is obtained. In the case of supervised tasks, such as ours, this predicted output is compared with the desired output to give an error, and this error is back-propagated across all the layers, to update the values (or weights) assigned to the connections between them. The input is then reevaluated with the new weights, and so on and so forth, improving the performance with every iteration as the network 'trains' itself. Once trained, the network should, given an input, be able to accurately predict the desired output.

Artificial Neural Networks are the foundation of deep learning. In fact, the term 'deep learning' comes from these 'deep' networks - referring to the fact that they contain multiple hidden layers between the input and the output, and can hence be termed deep. Deep Learning is a step closer to human-like artificial intelligence in comparison to traditional Machine Learning - it attempts to replicate the working of the human brain. Deep learning networks are able to learn features themselves, and solve problems like complex input-output mappings on their own. However, they require a lot of data, and hence computing power, in order to learn the general and correct solution to problems. Their performance generally improves as we provide them with more data.

There are various different kinds of neural networks - we have used two in our project: Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) Network. They are briefly explained as follows:

A. Convolutional Neural Network

These are used primarily in image-processing and computer vision systems. They are based on the concept of *convolution* (Saha, 2018), a process through which relevant features can be extracted from the images. The image/video input can be represented as a multi-layered collection of 2-dimensional grids (or matrix) of pixels. By performing a convolution on each of these individual matrices, we obtain smaller result matrices, called features. These features are then used for the purpose of classification.

Although this basic structure is extremely well-suited for extracting features from images, it can be applied to text data as well (Britz, 2015). In order to do this, we must organize the text in a similar matrix, or grid, format of numerical values.

Figure 8: A Sample Text Matrix

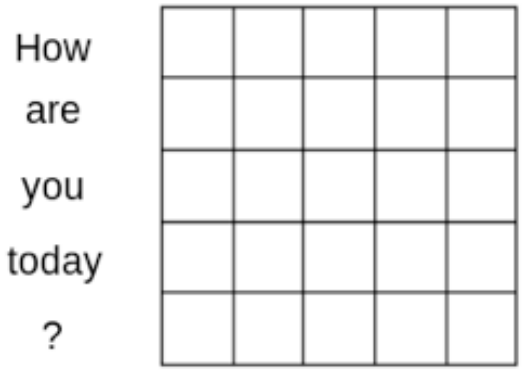


Figure 8 shows us the structure of such a grid for a simple sentence: *How are you today?*. Each word is encoded as a single row (a unique vector), and each cell within every row will have a numerical value. But how do we convert words to numerical vectors? That is where **word embeddings** come in.

1. Word Embeddings

Word embeddings are vector, or numerical, representations of words of human language, where each word is represented by a unique vector. Thus, we can treat them as an encoded form of words that our neural networks are able to understand. There are multiple ways of generating word embeddings - asking our neural networks to generate them from our text vocabulary is one. Another option is using pre-trained word embedding models such as Google's word2vec , one that provides a better representation of words, as words with similar meanings, such as 'hello' and 'hi' are encoded as vectors that are very close together mathematically. Words like 'hello' and 'stone', for instance, having no connection between them, would be far apart. Thus, they enable us to capture the meaning of words better.

Google's BERT embedding model computes these word embeddings in a different way, and aims to capture not just the meaning of the words but also their context in a sentence. Thus, the word *break* would have a completely different embedding vector representation in the sentences: "Give me a break", and "Do not break the glass".

We have used three different forms of word embeddings in our work, as described earlier in Section IV-C1.

B. LSTM Network

Long Short Term Memory Networks are a special class of **Recurrent Neural Networks** (RNNs). As the name suggests, Recurrent Neural Networks are neural networks that allow their output to recur through loops in the neurons (or nodes), i.e. the previous information persists in the network and is used for computing future outputs. This behaviour aims at mimicking the human memory, where we draw understanding of current events, such as the climax of a movie, from past ones, such as the events that preceded it (Olah, 2015). The problem with traditional RNNs is that they do not have a long memory - they cannot capture long-term dependencies between words as described in Section IV-C2. LSTMs aim to solve this problem by maintaining a 'cell state' in every neuron. This cell state remembers information deemed to be important across multiple iterations, and thus allows the network to remember relevant context, even with long-term dependencies and relationships.

Electrodermal Responses: When is Bigger Really Better?

Donald J. Krapohl[1]

Introduction

Manual scoring of polygraph charts has been accepted practice since the 1960s. Among the traditional polygraph channels, the responses in the electrodermal channel has repeatedly been shown to be most closely correlated with ground truth, making it responsible for roughly half the information in polygraph charts. Scoring the electrodermal channel is also the easiest. With 7-position scoring and its variants, scorers look for differences between the amplitude of an electrodermal response (EDR) to relevant and comparison questions. The greater of the two amplitudes will determine whether the score is in the positive direction (toward truthfulness) or negative direction (toward deceptiveness). Over the course of multiple question presentations, EDR scores and those from other polygraph channels are tallied. Decision rules are based on those tallies.

There are some scoring rule dissimilarities among systems. Speaking specifically to the EDA, there does not appear to be uniform agreement as to the minimum degree of difference required between EDR amplitudes before scores can be assigned. EDR scores, and by extension the tallies on which polygraph outcomes are based, are directly affected by which minimum is used. Cleve Backster was the first to propose minimum differences between two EDR amplitudes for assigning a score. According to Richard Weaver (1980), Backster initially required a difference of 3:1 or more to assign a score, though he modified it to 2:1 in 1976. The US Army Military Police School (early forerunner of the National Center for Credibility Assessment, or NCCA) and the University of Utah scoring systems similarly required a difference of at least 2:1 between the amplitudes of two EDRs to give a score different from 0. We were unable to locate any research indicating the 2:1 ratio had been developed empirically. The historical basis suggests the 2:1 ratio was simply a convenient and useful heuristic.

The NCCA changed from the 2:1 requirement sometime before 1999. It adopted instead a standard that any noticeable difference in amplitude between two EDRs was enough to give a non-zero score (Swinford, 1999). It is known as the "Bigger-Is-Better" Rule (BIBR). The rule is described as the following in the most recent publicly available NCCA pamphlet on manual scoring (2012):

> 5.10. Bigger-is-Better Principle. How do you evaluate two comparative responses, irrespective of whether they are similar or dissimilar in nature, where the amplitude ratio is less than two-to-one? The principle of "bigger-is-better" was adopted to address this situation. When the ratio between comparative responses is less than 2:1, the response with the more significant amplitude will receive the value. (page 31).

[1]The author is APA Past President and regular contributor to this publication. Questions and comments can be directed to APAkrapohl@gmail.com. The views expressed in this paper are solely those of the author.

The author is very grateful for the thoughtful suggestions and comments for an earlier draft of the paper by APA Editor Mark Handler and a blind reviewer.

NCCA scoring policies are used by all federal polygraph programs and a majority of state and local law enforcement agencies. The BIBR is also taught in most polygraph schools. It is likely the NCCA scoring methods are used by many or most field polygraph examiners. What has yet to appear in any publication we could find is how much bigger one reaction must be to give a score. Although the expression "more significant amplitude" in the NCCA scoring pamphlet is useful, it is also imprecise and subject to interpretation. How much is needed to be significant? Jimmie Swinford, who taught chart interpretation at the government polygraph school in that era, offered to define it better during his instruction. In 2008 he gave a presentation to the Indiana Polygraph Association in which he defined it as a "visually discernable difference" (slide 199). That is, if a difference in EDR amplitude between to questions can be seen by the scorer the larger reaction can receive a score.

Seeing amplitude differences might be straightforward when they are substantial. It may be less so as the amplitudes begin to approach one another in size. Different scorers assessing the significance of two very similar amplitudes may be influenced by several factors, including experience and training. Their judgment may also be affected by how the data are displayed. It has been the experience of the writer during presentations on scoring that experienced examiners are more reluctant to assign scores when the gain setting produces smaller amplitudes than when the amplitudes have been magnified by a higher setting of the same data. "Visually discernable," it seems, may be in the eye of the beholder.

A first approximation for the effect of the BIBR might be possible by examining EDR scores of experienced scorers conducting blind analysis on the same set of polygraph charts. If individual discretion is permitted in determining what constitutes a "more significant amplitude" it can be expected that score assignment would vary among scorers. The degree of variability will provide a rough index of how much they disagree on what constitutes a "more significant amplitude." As a first look we determined to test how close experienced blind scorers get to 100% agreement on whether to assign non-zero EDR scores when the BIBR is the scoring rule.

We had at our disposal the score sheets of the three US government polygraph examiners who scored the cases in the Blackwell study (1998). There were 100 confirmed field cases, 65 deceptive and 35 truthful. The scorers used the 7-position scoring system and among the 100 cases there were 861 opportunities to assign EDR scores. To assess the degree of reliability among scorers who used the BIBR we were only interested in the question of whether the examiners decided to assign non-zero scores. Therefore, we collapsed the examiner EDR scores to simply +1, 0 and -1 and then determine how often pairs of scorers agreed on each EDR score assignment. Chance agreement between pairs of scorers on whether they assigned any of these three scores was 0.33. When comparing all possible combinations of scorers we found the average proportion of agreement between pairs of scorers was 0.79. While impressive and greater than chance ($z = 19.1$, $p < .05$), a strict criterion for when to assign non-zero scores could have hypothetically achieved up to 100% agreement. Validity cannot exceed reliability, and the Blackwell data make the case for more stringent criteria for score assignment than the more general BIBR.

A new way of looking at EDR score assignment is to consider the effect of a rule set on whether the tally of scores for each case ends up on the correct side of zero. The more often a rule set produces a higher proportion of negative sums of scores for deceptive cases and positive score tallies for truthful cases, the higher will be the performance of that rule set. It is not a matter of simply assigning more scores, but rather maximizing the assignment of correct scores while minimizing those that can contribute to decision errors.

To begin there must be an objective and rigid rule set. Then one can systematically vary the criteria in the rule set to determine the peak

performance where the proportion of correct decisions has been maximized.

We came to appreciate the advantages of the fix-ratio approach introduced by Backster, if not the ratios he advocated. Backster's approach is convenient and well socialized in the polygraph profession. Because it relies on measurements of amplitudes it is objective, perfectly so if automated tools make those measurements. We set about using automated measurements and ratio calculations to determine whether there is an optimal ratio of EDR amplitudes where the highest proportion of total scores are in the correct direction.

Method

Data

Objective measures of EDR amplitude were taken from the data used to develop OSS-2 (Krapohl, 2002). The file consisted of 300 confirmed cases, half of them deceptive. The technique was the Federal Zone Comparison Technique (FZCT). Each case had three relevant questions and three charts. The FZCT scoring rules normally allow examiners to score against the stronger reaction when a relevant question is bracketed between two comparison questions. However, only one of the three relevant question in the FZCT is so bracketed, and the rules do not permit relevant question rotation within the test. Therefore, the optimal EDR ratio for one test question may be different from the optimal ratio for the other two questions that are not bracketed by comparison questions. For this reason, we only compared the EDR from each relevant question to the immediately preceding EDR from the comparison question. We created ratios by dividing the amplitude of the EDR at the relevant question by the amplitude of the comparison question EDR.

Procedure

We were interested in how changing the minimum ratio for score assignment would affect total EDR scores. More specifically, we wished to determine whether a particular minimum ratio could result in more cases having total EDR scores in the correct direction than other minimum ratios.

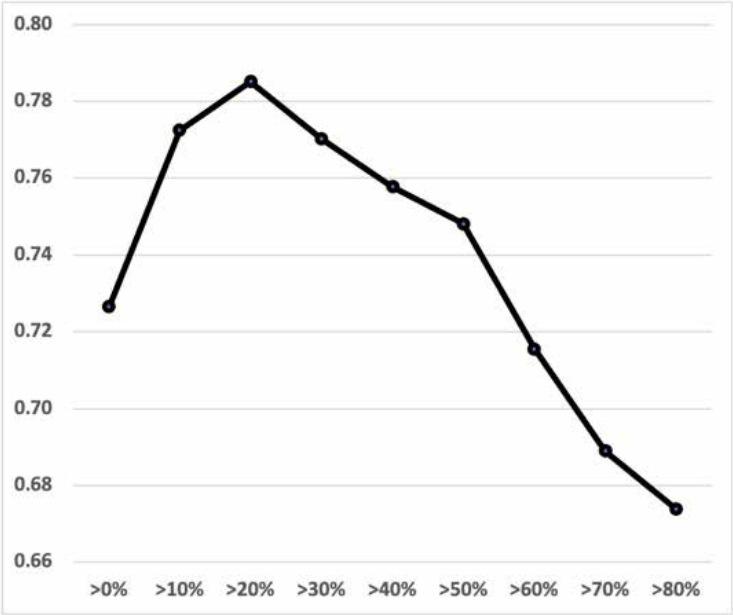
To assess the effect of different minima, we systematically increased the minimum ratio for score assignment from 1.0:1 to 1.8:1 in 0.10 increments. The smallest ratio was any score in which one EDR was bigger than the EDR against which it was being scored (>1.0:1). It did not matter how much bigger. If either EDR compared to the other exceeded the ratio of 1.0:1 a score was assigned. A positive score was given when the EDR at the comparison question was larger, and a negative score if the EDR at the relevant question was the greater. We repeated these steps for differences of 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80%, corresponding with ratios 1.1:1 through 1.8:1, respectively, resulting in nine different minimum ratios.

In the next step we calculated a detection efficiency coefficient (DEC; Kircher, Horowitz & Raskin, 1988) for the decisions resulting from the nine minimum ratios. The DEC represents the strength of a relationship between test results and ground truth. For test results, total scores by case were coded as +1 if they were greater than 0, -1 if less than 0, and 0 if the total was exactly 0. Confirmed truthful cases were coded as +1 and deceptive cases as -1. A point bi-serial correlation test was performed on the nine sets of data, producing the DEC for each of the sets. We then plotted the correlation coefficients by minimum ratio differences. A DEC of 0.0 would indicate there was no relationship between total EDR scores and ground truth, and a DEC of 1.0 would represent a perfect correlation.

Results

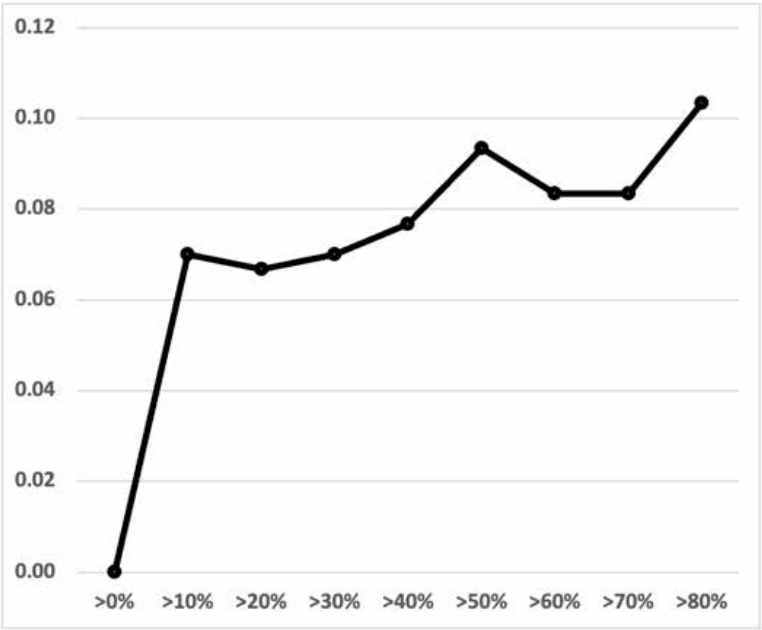
Changing the minimum ratio for score assignment influenced the detection efficiency coefficient. We plotted the DECs across the nine ratio differences, from >0% to >80% difference. Figure 1 shows the results. As can be seen, the largest change occurs between the minima of >0% and >10%.

Figure 1. Detection efficiency coefficients between ground truth and EDR scores at escalating minimum ratios between >0% and >80% in 10% increments for 300 confirmed cases.



The proportion of Inconclusive cases was also related to the minimum EDR ratios. When any difference in EDR amplitudes was considered sufficient to assign a score, there were no cases which had total EDR scores of 0. This is unsurprising. At the largest minimum ratio tested the proportion of cases having a total EDR score of 0 increased to 0.103.

Figure 2. Proportion of cases with EDR scores summing to 0 at escalating minimum ratios between >0% and >80% in 10% increments for 300 confirmed cases.



Discussion

The present data show the BIBR rule works quite well. The data also suggest the BIBR might be optimized. When considering the trend in Figure 1, the DEC shows its largest improvement when changing from a minimum ratio of >0% to one of >10%, going from a DEC of 0.727 to a DEC of 0.772. The DEC peaks at the >20% minimum ratio, with a DEC of 0.785, and falling off thereafter. These data suggest that any minimum difference between 10% and 50% will, in the long run, outperform a minimum ratio of >0%. The greatest performance is between >10% and >20%. These findings map closely to Handler et al. (2010) who reported diagnostic efficiency was highest at a minimum EDR ratio of about 10%.

Rates in which total EDR scores were 0 also varied across the minimum ratio differences used to assign scores. A total EDR score of 0 means, in practical terms, the EDA data were neither in the correct nor incorrect direction. Polygraph decisions in those cases would wholly rely upon the scores of the remaining data channels. In the present study the proportion of cases in which EDR scores summed to 0 was maximal (0.10) when there was a requirement for an 80% difference or greater to assign an EDR score. This dropped to 0.00 when score assignment required only one EDR to be larger than the other. As with the DEC trend, the greatest change took place between the minima of >0% and >10% where the proportion of total scores of 0 increased from 0.00 to 0.07.

We return now to the question we posed in the title; When is bigger really better? A generic answer is that, according to our data, bigger is always better. Within the limits of the ratios we examined the data show that any minimum ratio one chooses will produce significant detection efficiency coefficients. Strictly speaking, there does not appear to be any ratio that is wrong, though it is also clear that not all minimum ratios perform equally.

A post hoc question might be this; When is bigger the best? Our data point to a minimum ratio from 10% to 20% between two EDRs for assigning a score. It is between these two ratios the DEC has its maximum values. The data suggest scorers who use minimum EDR amplitude differences between 10% and 20% will, over a large number of cases like those used in our sample, obtain the most diagnostic information available in EDR amplitudes.

Limitations

Our findings are more germane to 3-position scoring systems, including the Empirical Scoring System, than to others such as Rank Order or 7-position which were not assessed in this paper.

The study also looked only at the effect on single-issue examinations. Mixed-issue examinations, where decisions are based exclusively on sums for individual questions, are expected to show more variability because there are fewer samples to consider. With greater variability it is likely the optimal EDR ratios for multiple-issue testing will be different, perhaps higher than they are for single-issue testing.

Finally, the ratios developed here relied on scoring each relevant question to the immediately preceding comparison question. Testing formats in which scorers can score to the stronger of two comparison questions may find a different optimal ratio for score assignment, possibly higher.

References

Blackwell, N.J. (1998, Sep). *PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy when Scoring Examination from Actual Criminal Investigations.* Department of Defense Polygraph Institute, DoDPI96 – P- 0001. Ft. McClellan, AL. Published in Polygraph (1999), 28(2), 149 – 175.

Handler, M., Nelson, R., Krapohl, D. and Honts, C.R. (2010). An EDA primer for polygraph examiners. *Polygraph, 39*(2), 68 – 108.

Kircher, J.C., Horowitz, S.W, and Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior, 12*(1), 79 – 90.

Krapohl, D.J. (2002). Short report: Update for the Objective Scoring System. *Polygraph, 31*(4), 298 – 302.

Krapohl, D.J., & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. *Polygraph, 35*(1), 55-63.

National Center for Credibility Assessment (2012, Jan). *Test Data Analysis: Numerical Evaluation Scoring System Pamphlet.* Ft. Jackson, SC.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph, 28(*1), 10 – 27.

Swinford, J. (2006, Jun 27). *Federal Test Data Analysis Procedures.* Presentation to the Indiana Polygraph Association.

Weaver, R.S. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. *Polygraph, 9*(2), 94 – 108.

Bigger is Better for Automated Scoring: Analysis of Minimum Constraints for RQ/CQ Ratios

Raymond Nelson

Abstract

An archival sample of n=300 confirmed field polygraph examinations was used to study the effects of minimum constraint ratios, from 1:1 to 2:1 in increments of .05, for automated feature extraction and automated score assignment. For respiration data 95% of the scores were zero (0) at a minimum constraint ratio of 1.6:1. In contrast, approximately 55% of the EDA scores were non-zero and 39% of the cardio scores were non-zero at the same (1.6:1) ratio. LogRC Ratios were optimal with no minimum constraint, indicating that automated scoring methods are reasonable when they attempt to make use of any measurable difference that can be extracted from relevant and comparison questions. For signed integer scores, the correlation coefficient (similar to DEC) for the 2700 numerical scores was largely unaffected by any constraint for cardio data. The correlation coefficient for numerical scores of EDA data was minimally affected by the series of constraints, beginning at .425 at 1.05:1, then rising slightly to .450 at ratio of 1.2:1, and ending at .385 at the maximum constraint ratio of 2:1. Score correlations for respiration scores, together with the aggregated score correlation (shown in orange), suggest that constraining the respiration score extraction to the range from 1.2:1 and 1.6:1 may be useful to optimize the contribution of respiration scores to correct vs incorrect conclusions. Data from this analysis indicate that no minimum constraint ratio is needed for automated analysis methods for EDA or cardio and provide general support for the validity of the bigger-is-better rule.

Introduction

All scientific conclusions, in both scientific research and scientific testing, are made with regard to other possible conclusions. The process of science is intended to evaluate the strength of available evidence to support each of the different possible conclusions that attempt to answer basic questions about the universe and reality. What is it? How does it work? Why? Regardless of whether science occurs at the level of theoretical physics or at the level of practical forensic and risk-management decisions about how best to proceed with a single individual, reproducibility of analytic results has become a de facto standard or expectation for all areas of scientific research and scientific testing (Peng, 2011). The need for reproducibility can be easily observed in credibility assessment testing – beginning with the fact that test data are permanently recorded. An ability to record data is foundational to an ability to study the signals in different ways so that analytic methods can be optimized.

Advancements in technology during the early history of the polygraph profession involved both the development of sensors that can provide access to physiological signals that are correlated with deception and truth-telling, and methods to record changes in physiology so that they may be studied more carefully and studied repeatedly. Today we know that although deception itself cannot be measured physically, we are not likely to ever find any physiological activity that is uniquely associated deception. Instead, all polygraph signals will most likely continue to involve the autonomic nervous system and multiple aspects of the cerebral cortex. And all physiological activity will likely continue to be associated with multiple types of human behavior. In short: there is no such thing as "Pinocchio's nose."

The strength of correlation of different physio-

logical signals, along with the degree to which different signals may covary, will remain an underlying concern to anyone involved in polygraph validity research (or discussions about polygraph validity). To be use useful, physiological signals will ideally correlate with the criterion at a statistically significant level, but will not covary so strongly that they are redundant. Useful signals will contribute unique, non-redundant, information to a structural model (i.e., a mathematical/statistical representation of the phenomena of interest). Redundancy of physiological signal can be observed when adding additional data to the model does not increase the effectiveness of the model, even though the added information is known to be correlated with the phenomena of interest. In other words, the simplistic adage "more information is always better" is untrue: more information is better when it increases the effect size of interest. If the added information does not increase the effect size of interest the actual effect will be an increase in risk for confusion and unreliability. The result of all of this is that scientific tests are often constructed of signals of moderate correlation strength – because signals for which the criterion correlation is strong will tend to covary so much they can become redundant.

During the early part of the 20th century the kymograph was the best available technology to record polygraph data for subsequent analysis, and re-analysis. Analytic methods through the mid-century period relied almost uniformly on the un-quantified intuition and experience of the expert observer. Over time, toward the latter half of the 20th century, the need for improved consistency and skill development among a variety of experts led to an emphasis on numerical scoring systems such as the seven-position system and three-position scoring method. Towards the latter half of the 20th century we saw an exponential increase in the availability computing technology. Powerful and (relatively) inexpensive computing technology has influenced virtually every aspect of social and professional life – including recreation, entertainment, communication, transportation, education, administration, employment, news and information, publication, and even science and scientific testing.

Today – well into the 21st century – there is no area of society and no area of science that does not make use of computing technology to record and analyze data. The kymograph of the early 20th century is today virtually completely supplanted, in both polygraph field practice and polygraph research, with analog to digital converters and computerized encoding systems that record polygraph data not as a tracing on a cylinder or paper scroll but as a time-series of recorded numbers stored on an electronic media. Data are processed for display in the familiar form of time-series tracings on a computer screen. A convenient aspect of all of this is that older polygraph examiners can plot or print their "charts" onto paper and inspect them visually in ways similar to what they have done in decades past.

Whereas the factors that influence the plotted lines in the days of early polygraph instrumentation were entirely mechanical – involving the moving mass of carefully engineered hardware, including the friction coefficients of pivots and bearings along the myriad of adjustments and calibrations necessary to ensure that recorded data, encoded as ink on paper, would be useful – polygraph signal processing can today be more carefully and precisely designed through the careful efforts of electrical engineers who understand our hardware requirements and through software engineers and data scientists who can enable us to make use of digital signal processing methods and statistical methods with more power than those we used during the era when all computations were done manually.

Electronic engineering and digital signal processing methods can provide far greater precision and reliability, and with much greater convenience and economy, than mechanical solutions of the past. In contrast to mechanical polygraph systems, in which filtering and smoothing was sometimes an unintended or unanticipated byproduct of the friction of the weight of the capillary ink pen on the scrolled paper, computerized polygraph systems of today – with high sampling rates and high resolution analog-to-digital conversion – can provide data that is of higher fidelity, in terms of recording and representing physiological activity, than ever in the past.

Simultaneous with advances in polygraph testing methods, signal processing and data recording, data analysis methods have also advanced as a result of available computing technologies. Polygraph professionals now have access to both empirical reference distributions (Krapohl & McManus, 1999; Krapohl, 2002; Nelson, Krapohl & Handler, 2008; Nelson & Handler, 2015) and multinomial reference distributions (Nelson, 2017; 2018). The availability of computer-based statistical reference models has led to the potential for convenient application of both frequentist and Bayesian statistical methods in polygraph field practice.

Today, in the 21st century, we have the capability for both digital recording of polygraph signals and the potential convenient use of powerful mathematical and statistical methods that can go well beyond what polygraph professionals are willing to attempt with pencil and paper. We also have the capability for automated feature extraction – and this will be inherently more reliable than feature extraction through visual pattern recognition methods that may have been the best available solution for analog polygraph instruments. Deception and truth-telling are complex problems – beginning with the complex asymmetry of even achieving a completely satisfactory epistemological/philosophical definition of deception and truth. It is also not surprising that the analysis of credibility assessment test data is inherently complex – and therefore subject to a variety of forms of bias, subjectivity and inconsistency.

The magnitude of complexity surrounding polygraph feature extraction becomes quickly apparent when considering the combination or interaction of factors that can influence a numerical score, including feature extraction at both the relevant-question (RQ) and comparison-question (CQ) along with the comparison of these two values. The most realistic solution for the future of the polygraph profession will be to harness the power of digital computers to record process and analyze the variety of complexities and interactions, including the task of automated feature extraction and automated score assignment. To do otherwise – to limit polygraph methodology to mid-century methods from the pre-computer epoch – will be to invite eventual disruption. Fortunately, a great deal of knowledge and methodology exists for this purpose. Computing power and analysis tools today are abundant and inexpensive – quite often they are free and open source. This project is an optimization study of automated numerical score assignment as a function of the ratio of RQ and CQ and pairs. The question of interest is whether there exists a set of minimum RQ/CQ ratio constraints that will maximize the diagnostic information that is achieved in the numerical scores for each of the polygraph recording sensors.

Methods

Data

Data for this project were n=300 confirmed field exams that were conducted using the Federal Zone Comparison Test (FZCT, Department of Defense, 2006) format. Sample cases were conducted by a variety of federal, state, and municipal law enforcement agency and were subsequently included in the confirmed case archive at the Department of Defense Polygraph Institute (now the National Center for Credibility Assessment). All cases consisted of three iterations of a question sequence that included three relevant-questions (RQs) and three comparison-questions (CQs) in addition to other procedural questions that are not subject to numerical or statistical analysis. All exams consisted of three completed test charts. [Refer to Nelson (2015) and Department of Defense (2006) for general information on the comparison question test and how the sample cases were conducted.]

All of the sample cases included the standard array of sensors, including upper and lower respiration sensors, an electrodermal activity sensor, and cardiovascular activity sensor from which responses would be extracted and numerical scores assigned. This sample was previously used in the development of the OSS-2 scoring method (Krapohl, 2002), at which time response features were extracted from the recorded data using a computer software program (Extract.exe, Harris, in Krapohl & McManus, 1999) that was developed to objectively extract the Kircher feature measurements from respiration, EDA and cardio data of computerized polygraph data. A total of 21600 measurements were available for the n=300 field cases with three iterations of

a question sequence that included three RQs and three CQs. Data were imported to the R Language for Statistical Computing (R Core Team, 2019) for analysis.

Analysis

All iterations of all relevant questions (RQs) were evaluated using the comparison question selected according to the standardized procedure for the FZCT format. The RQs are labeled R5, R7 and R10, while the comparison questions (CQs) are labelled C4, C6, and C9. For each sensor, the first RQ, R5, was evaluated with CQs that are immediately preceding and immediately following the RQ – either C4 or C6 on the first recorded chart, though the questions may be rotated for subsequent charts – depending on which CQ produced the greater change in physiological activity. The second and third RQs, R7, and R10 were evaluated with the preceding CQ – C6 for R7 and C9 for R10, though the order may be rotated for some recorded test charts. An RQ/CQ ratio., referred to as an RC Ratio, was calculated for each pair of questions. For EDA and cardio sensors greater extracted values indicate greater changes in physiology. In contrast, for the respiration sensor, smaller extracted values represent greater changes in physiological activity.

RC ratios will conform to an asymmetrical distribution, bounded by 0 and ∞ (infinity) with a mean of 1 and a potentially infinite range of values between 0 and 1, along with a potentially infinite range of values between 1 and infinity. When the RQ value was greater than the CQ value the RC Ratio was a value between 1 and infinity. When the CQ value was greater than the RQ value the RC Ration was a decimal value between 0 and 1. To avoid this asymmetry the natural logarithm was taken for each RC Ratio, referred to as a logRC Ratio, The resulting distribution of logRC Ratios was a symmetrical distribution with a mean of 0 and an infinite number of potential values between 0 and ∞ (infinity) along with an infinity number potential values between 0 and -∞ (negative infinity). RC Ratios between 0 and 1 produced negative logRC Ratio values between 1 and negative-infinity, while RC Ratios between 1 and infinity produced positive logRC Ratios between 0 and infinity. Table 1 shows an example of the use of the natural logarithm to produce ratios that are symmetrical around 0.

Table 1. Examples showing use of the natural logarithm to achieve a symmetrical distribution of logRC ratios.

	RQ Value	CQ Value	RC Ratio	logRC Ratio
Ex 1	300	200	1.5	0.4054651
Ex 2	200	300	.67	-0.4054651

Notice, in Table 1, how RC Ratios are not symmetrical around 1 while logRC Ratios are symmetrical around 0. This symmetry make it possible to make use of linear statistical calculations such as the correlation coefficient. Before proceeding further, it was necessary to adjust the sign values of the logRC Ratios for EDA and cardio data so that negative logRC Ratios correspond to deceptive scores while positive logRC Ratios correspond to truthful scores for all sensor, including the respiration, EDA and cardio.

Twenty-one thousand six-hundred (21600) measurements were taken from the n=300 cases, from which a total of 10800 logRC Ratios were calculated for the three iterations of three RQs, for the thoracic and abdominal respiration sensors, EDA sensor and cardio sensor for each of the n=300 sample cases. After combining the data for the thoracic and abdominal respiration sensors there were 8100 logRC Ratios, including 2700 values for each recording sensor: respiration, EDA, and cardio. To remain consistent with the familiar intuition for integer scores used in polygraph field practice, logRC Ratio of + sign value correspond to truth-telling while integer scores of – sign value correspond to deception. During the course of the analysis, automated signed

integer scores were assigned to the 10800 logRC Ratios using the bigger-is-better-rule (BIBR: National Center for Credibility Assessment, 2017). Numerical scores of this type are similar to the scores that human experts would assign using manual scoring methods such as the Federal three-position scoring method (National Center for Credibility Assessment, 2017) or the Empirical Scoring System (ESS: Nelson, Krapohl & Handler, 2008).

Respiration data

Thoracic and abdominal logRC Ratios were combined to a single vector of 5400 values and the point-biserial correlation for respiration scores was r_{pb}=.184. For the thoracic respiration sensor alone, the value was r_{pb}=.209, and for the abdominal sensor alone it was r_{pb}=.161. For the combined respiration sensor data, the maximum logRC Ratio was 3.8 (a ratio of 45:1). A maximum constraint value was applied iteratively from +/-2 to +/0 3.5 and was found to optimize the point-biserial correlations at +/-2.7 (a ratio of 14.9). That is, logRC Ratios were coerced to zero (0) if they exceeded the values 2.7 or -2.7. LogRC Ratios for respiration data were more likely to contribute to incorrect scores than to correct scores when they exceeded this level. Twenty-four (24) of the 5400 respiration scores (<0.5%) exceeded the maximum constraint. With the maximum constraint value, the point-biserial correlations were r_{pb}=.216 for the thoracic sensor and r_{pb}=.182 for the abdominal sensor. An outer or maximum constraint can improve the correlation coefficients for aggregated respiration scores. However, because the goal of this project was to study optimal minimum constraints no further optimization was performed on the outer constraint for respiration scores. The maximum constraint ratio was retained for the remainder of the analysis.

Respiration scores for the thoracic and abdominal sensors were then combined to a single set of 2700 scores using the procedure described by Nelson and Krapohl (2017). Using that procedure. The combined logRC Ratio was coerced to zero (0) if the sign values are opposite, and was set to the value with the greater absolute value if not opposite. After combining the two logRC Ratios to a single respiration score for each iteration of each RQ for each case the point-biserial correlation was r_{pb}=.211. Respiration data were also combined by averaging the logRC Ratios for the two respiration sensors. The point-biserial correlation for the averaging method of combining the sensor data was r_{pb}=.215, and exceeded that of the procedural method. For the combined respiration vector, five (5) of the 2700 respiration logRC Ratios were zero (0). Separate vectors of logRC Ratios were retained for analysis, including 2700 values for the thoracic sensor and 2700 values for the abdominal sensors. Thoracic and abdominal information would be combined in a later step for each iteration of a series of minimum ratio constraints.

The logRC Ratios were then standardized to a mean of zero (0) and standard deviation of one (1). There was no difference in the correlation (r_{pb}=.215) for the standardized logRC Ratios for the averaged thoracic and abdominal respiration sensors. Because standardization offered no advantage, the remainder of the analysis was completed with the un-standardized respiration data. Standardized values will have a common metric, with mean=0 and sd=1, and can be calculated at a later time when combining data from different sensors.

LogRC Ratios for combined respiration data were then aggregated by averaging the three iterations of the three RQs for each case. The point-biserial correlation for the mean logRC Ratios for respiration data was r_{pb}=.408. For the individual sensors the point-biserial correlation after aggregating the scores for each case was r_{pb}=.420 for the thoracic sensor and r_{pb}=.359 for the abdominal sensor with the outer constraint. Without the maximum constraint the correlations were r_{pb}=.401 for the thoracic and r_{pb}b=.317 for the abdominal. It is not surprising that the aggregated logRC Ratios for each case produced a substantially stronger correlation coefficient than the logRC Ratios for each presentation of each RQ. Aggregating data has the effect of improving the signal-to-noise ratio within the information extracted from the recorded data.

A series of minimum constraints was evaluated, from 1.05:1 to 2:1 in increments of .05. Ratios from 1.05 to 2 were transformed to their natural logarithms so that they could be applied symmetrically to the logRC Ratios which have + and – sign values similar to the intu-

ition that field polygraph examiners use for truthful and deceptive numerical scores. The series of possible constraints was applied iteratively to the sample of n=300 cases. For each iteration of the series of minimum constraints, logRC Ratios were coerced to zero if they did not exceed the constraint. Integer scores using the BIBR were also coerced to zero for values for which the logRC was coerced to zero. The constraint value was applied separately to the thoracic and abdominal sensor data before combining the sensor data using the method described by Nelson & Krapohl (2017). Log RC Ratios were aggregated by averaging all iterations of all RQs for each case, and the point-biserial correlation was then calculated for the case criterion states coded as [+1, -1].

LogRC Ratios for each presentation of each RQ were then transformed to signed integer scores for which the thoracic and abdominal scores were combined using the method described earlier. Sign values for both the logRC Ratios and the signed integer scores conform to the familiar intuition for sign scores used by polygraph field examiners. Scores of positive (+) value correspond to truth-telling, and scores of negative (-) sign value correspond to deception. For each iteration of the series of minimum constraint ratios the proportion of non-zero logRC Ratios was calculated – also the proportion of non-zero signed integer scores – for the 2700 scores after combining the thoracic and abdominal sensor data. The proportion of correct non-zero signed integer scores was calculated by comparing the integer scores with the case criterion state coded as [+, -]. Finally, the signed integer scores were summed for each case and the point-biserial correlation was calculated for the numerical scores and the criterion state, coded as [+1, -1] for each iteration of the minimum constraint ratio.

For each iteration of the minimum constraint ratio the correlation of the 2700 signed integer scores and the case criterion state was calculated using a procedure similar to the detection efficiency coefficient (DEC; Kircher, Horowitz, & Raskin, 1988). DEC is calculated as the Pearson correlation between integer score codes [+1, 0, -1] and the criterion state [+1, -1], and are informative because they represent strength of information about correct, incorrect, and inconclusive outcomes in a single correlation statistic. This application of the DEC differed from its normal use in that DEC was initially described for use with classifications made with numerical or statistical cutscores using aggregated scores for each case using a complete array of recording sensors; it is use here with the individual scores for a single senor and with no numerical cutscores. In this usage the DEC correlation can the thought of as a numerical score correlation; it provides a measurement of the strength of information from the numerical scores at each minimum constraint ratio.

EDA data

For EDA scores the point-biserial correlation for the 2700 logRC Ratios was r_{pb}=.433. The maximum logRC Ratio was 3.4, corresponding to a ratio of 30:1 where the CQ value exceeded the value of the RQ. The minimum logRC Ratio was -4.7, corresponding to a ratio of 110:1, where the RQ value exceeded the CQ value.

A maximum constraint ratio was applied iteratively from +/-2 to +/-20 and was found to maximize the point-biserial correlation with a maximum constraint ratio of 7:1 with r_{pb}=.439. EDA LogRC Ratios that exceeded this level were more likely to contribute to incorrect scores than to correct scores. The remainder of the analysis was completed with this maximum constraint ratio. Ninety (90) of the 2700 logRC Ratios (3.3%) exceeded this constraint value.

LogRC Ratios were then standardized to evaluate the effect on the correlation coefficients. Standardizing the logRC Ratios did not change the point-biserial correlations. Because this project did not involve the aggregation of data between sensors, the remainder of the analysis was completed with the un-standardized logRC Ratios. Standardization will be advantageous, and can be accomplished at a later

After aggregating the logRC Ratios for all iterations of all RQs for each of the n=300 sample cases the point-biserial correlation was r_{pb}=.751. Aggregating data has the effect of improving the signal to noise ratio, and it is

therefore not surprising that the correlation for aggregated logRC Ratios exceeded the correlation for the individual logRC Ratios.

A series of minimum constraint ratios was evaluated from 1.05:1 to 2:1 in increments of .05. Results for the EDA data were re-calculated for each iteration of the constraint, including the point-biserial correlation of the aggregated logRC Ratios with the criterion state of each of the n=300 sample cases. For each iteration of the series of minimum constraints, logRC Ratios were coerced to zero if they did not exceed the constraint. The 2700 logRC Ratios were also transformed to signed integer scores [+1, 0, -1] and the criterion correlation with the sign scores was calculated for each iteration of the minimum constraint ratio. Results were also calculated for the proportion of correct sign scores and the proportion of non-zero scores. Finally, the signed integer scores were summed for each of the n=300 cases and the correlation was calculated of the aggregated integer scores with the case criterion state.

Cardio data

For cardio scores the point-biserial correlation for the 2700 logRC Ratios was r_{pb}=.179. The maximum logRC Ratio was 1.9 which corresponds to a ratio of 6.7:1. The minimum logRC Ratio was -3.9, corresponding to a ratio of 52:1.

A maximum constraint ratio was applied iteratively to the cardio data from +/-2 to +/-20 and was found to optimize the point-biserial correlation with a maximum constraint ratio of 12:1 with rpb=.180. LogRC Ratios that exceeded this the 12:1 level were more likely to contribute to incorrect cardio scores than to correct scores. The remainder of the analysis was completed with this maximum constraint ratio. Two (2) of the 2700 logRC Ratios exceeded the 12:1 constraint value.

LogRC Ratios were then standardized to evaluate the effect on the correlation coefficients. Standardizing the logRC Ratios did not change the point-biserial correlations. Because this project did not involve the aggregation of data between sensors, analysis further analysis of the cardio data was completed with the un-standardized logRC Ratios. Standardization of the cardio can be accomplished at a later time when data are combined for the array of recording sensors.

After aggregating the logRC Ratios for all iterations of all RQs for each of the n=300 sample cases the point-biserial correlation was r_{pb}=.460. Aggregating the cardio data has the effect of improving the signal to noise ratio, and it is therefore not surprising that the correlation for aggregated logRC Ratios exceeded the correlation for the individual logRC Ratios.

The same series of minimum constraint ratios, from 1.05:1 to 2:1 in increments of .05, was applied to the cardio data. Results were re-calculated for each iteration of the constraint, including the point-biserial correlation of the aggregated logRC Ratios with the criterion state of each of the n=300 sample cases. For each iteration of the series of minimum constraints, logRC Ratios were coerced to zero if they did not exceed the constraint. The 2700 logRC Ratios were also transformed to signed integer scores [+1, 0, -1] and the criterion correlation with the sign scores was calculated. Results were also calculated for the proportion of correct sign scores and the proportion of non-zero scores for the cardio data. Finally, the signed integer scores were summed for each of the n=300 cases and the correlation was calculated of the aggregated cardio integer scores with the case criterion state.

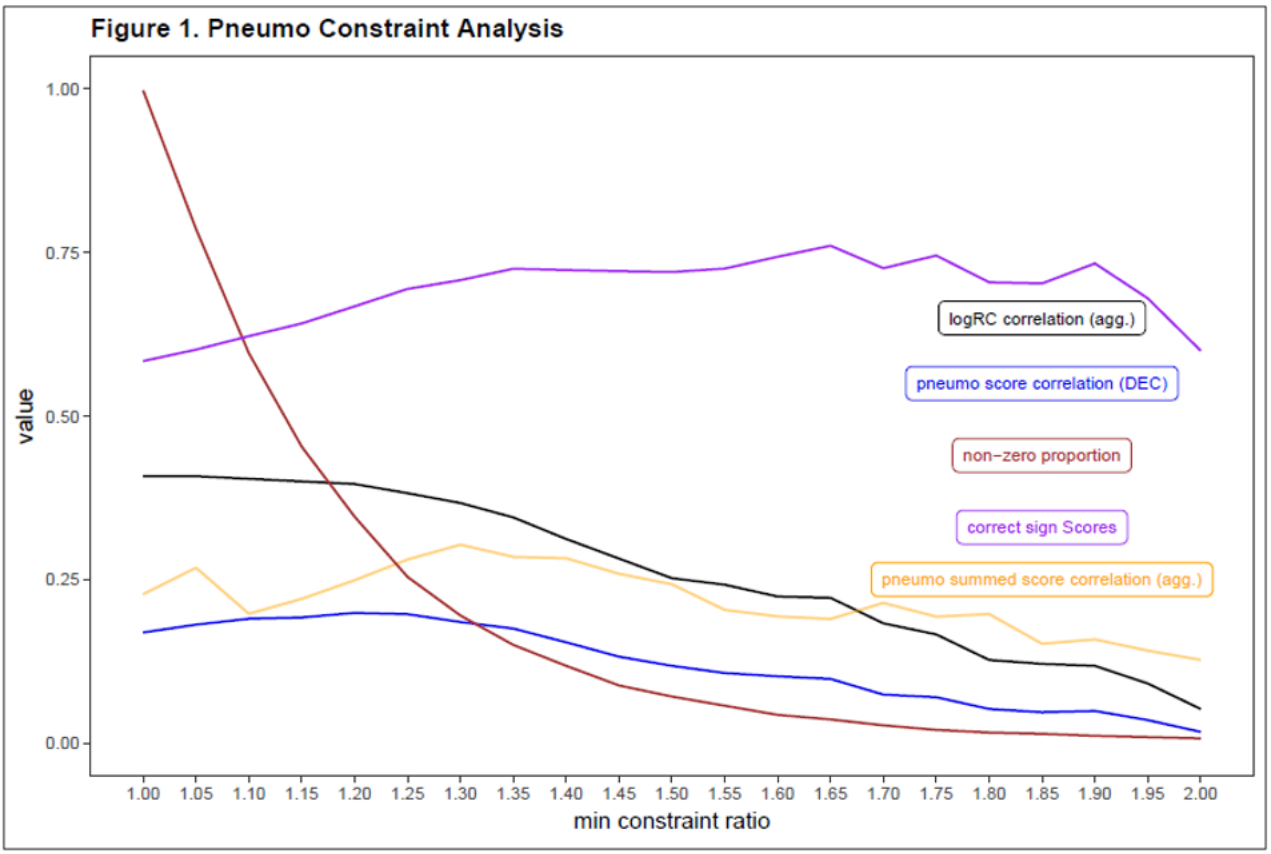
Results

Table 2 shows the mean point-biserial correlation coefficients for the 2700 logRC Ratios and the criterion state for each recording sensor, along with the mean point-biserial correlation coefficient for the aggregated logRC Ratios and summed integer scores. It can be seen in Table 2 that correlations for aggregated scores exceed those of the individual scores. This is an example of the value of using polygraph test formats with multiple RQs and multiple iterations of the question sequence.

Table 2. Point-biserial correlations for logRC Ratios, aggregated logRC Ratio, and summed integer scores.

	logRC Ratios (2700 scores)	Aggregated logRC Ratios (n=300)	Summed Integer Scores (n=300)
Respiration	UP=.216, LP=.182, comb.=.211	UP=.420, LP=.359, comb=.401	.228
EDA	.439	.751	.749
Cardio	.183	.460	.410

Figure 1 shows a plot of the series of minimum ratio constraints, from 1.05:1 to 2:1, with the respiration data, including the point-biserial correlation for the logRC Ratios and numerical scores after aggregating the data for each case. Also shown is the point-biserial correlation for the 2700 integer scores, along with proportion of non-zero scores at each minimum constraint level and the proportion of correct non-zero scores. Figure 2 shows a plot of the same information for the EDA data. Figure 3 shows the same information for the cardio data.



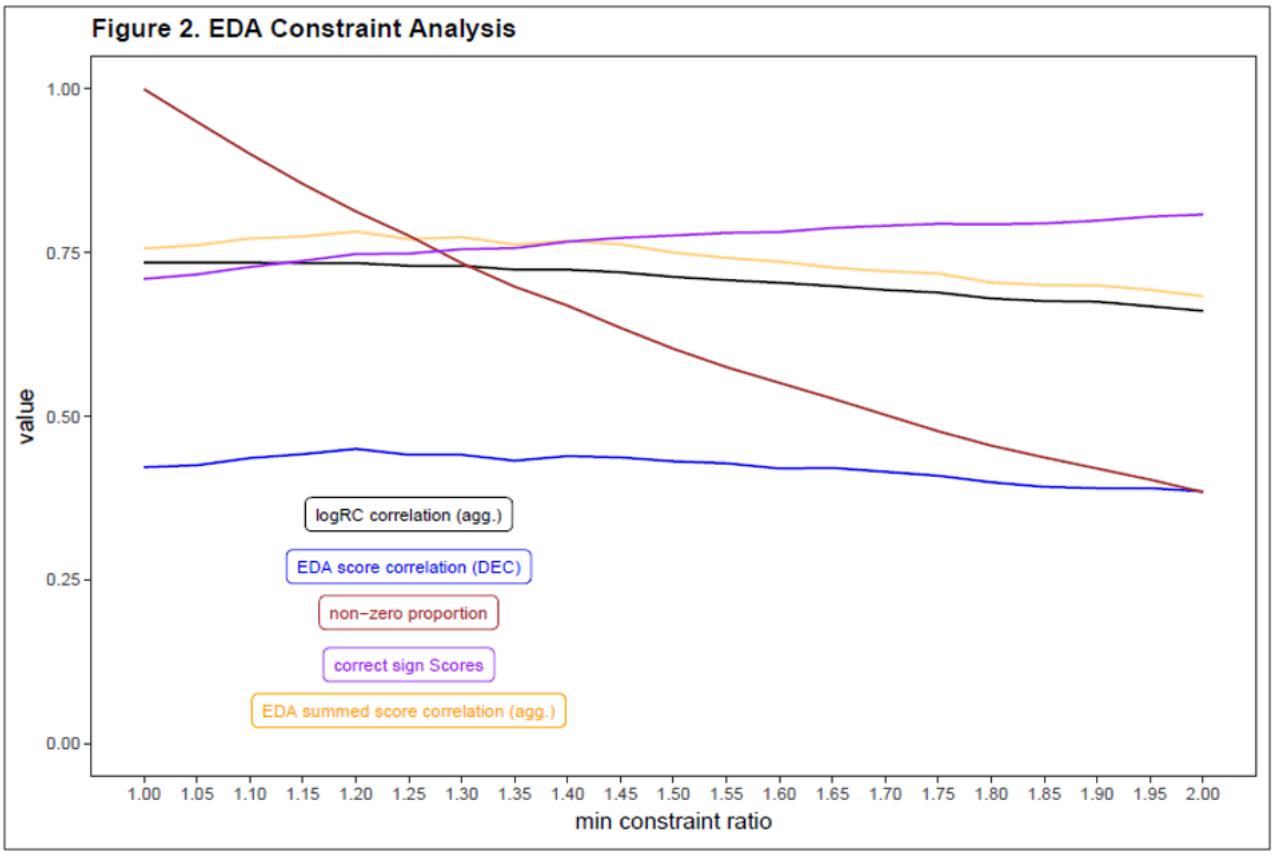


Figure 2. EDA Constraint Analysis
value
1.00
0.75
0.50
0.25
0.00
logRC correlation (agg.)
EDA score correlation (DEC)
non−zero proportion
correct sign Scores
EDA summed score correlation (agg.)
1.00 1.05 1.10 1.15 1.20 1.25 1.30 1.35 1.40 1.45 1.50 1.55 1.60 1.65 1.70 1.75 1.80 1.85 1.90 1.95 2.00
min constraint ratio

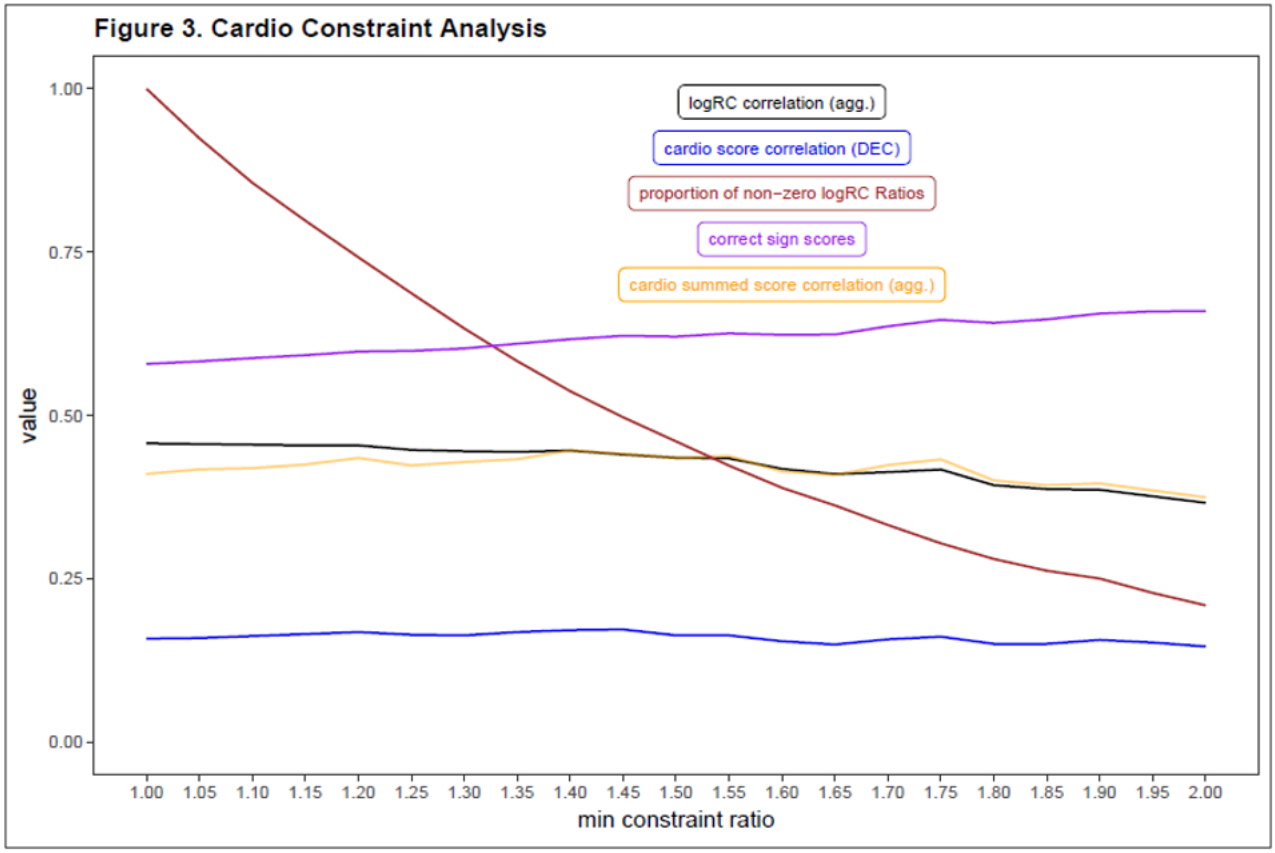


Figure 3. Cardio Constraint Analysis
value
1.00
0.75
0.50
0.25
0.00
logRC correlation (agg.)
cardio score correlation (DEC)
proportion of non−zero logRC Ratios
correct sign scores
cardio summed score correlation (agg.)
1.00 1.05 1.10 1.15 1.20 1.25 1.30 1.35 1.40 1.45 1.50 1.55 1.60 1.65 1.70 1.75 1.80 1.85 1.90 1.95 2.00
min constraint ratio

Discussion

For each of the recording sensors, respiration, EDA, and cardio, the point-biserial correlation between the aggregated logRC Ratios was greatest with no minimum constraint. Application of the series of increasing minimum constraints resulted in continuous weakening of the point-biserial coefficient. Not surprisingly, for each of the recording sensors the proportion of non-zero values – both logRC Ratios and numerical scores – was greatest with no minimum constraint and became progressively smaller as the minimum constraint ratio increase. This effect was most pronounced for the respiration data, for which 95% of the scores were zero (0) at a minimum constraint ratio of 1.6:1. In contrast, approximately 55% of the EDA scores were non-zero and 39% of the cardio scores were non-zero at the same (1.6:1) ratio. The correlation coefficient (similar to DEC) for the 2700 numerical scores was largely unaffected for cardio data. The correlation coefficient for EDA scores was minimally affected by the series of constraints, beginning at .411 and rising slightly to .450 at ratio of 1.2:1, and ending at .385 at the maximum constraint ratio of 2:1.

The proportion of correct signed scores – both logRC Ratios and numerical scores – increased to small degree across the range of increasing minimum constraint ratios for the EDA and cardio data. However, the magnitude of this increase was substantially less than the increase in the number of scores of zero (0) for these sensors. For respiration data, the proportion of correct signed scores increased to a peak at a ratio of 1.35:1 and became unstable a ratios exceeding 1.6:1. This instability can be attributed to the small number of non-zero-scores that remained at constraint ratios of 1.6:1 and higher.

The point-biserial correlation for the logRC Ratios (shown in black in Figures 1, 2, and 3), together with the score correlation (shown in blue) – similar to a DEC correlation for result, but calculated in this analysis with no numerical cutscores – provides a convenient synthesis of the complex information contained in this analysis. This correlations captures information about correct, incorrect and null (0) scores in a single numerical index for which the familiar intuition for correlation coefficients can be applied. For cardio scores no minimum constraint can be identified that will increase the effectiveness of the scores that can be extracted from recorded data. For EDA scores the effect of a minimum constraint ratio to improve the correlation coefficient of numerical scores was minimal. What remains is whether any statistically significant advantage exists for the use of a minimum constraint ratio for numerical scores. However, these data suggest that there is no advantage to the use of a minimum constraint with the logRC Ratios used in automated scoring methods – and this same conclusion would be observed using ratios without a log transformation. Score correlations for respiration scores, together with the aggregated score correlation (shown in orange), suggest that constraining the score extraction to the range from 1.2:1 and 1.6:1 may be useful to optimize the contribution of respiration scores to correct vs incorrect conclusions.

One obvious limitation of this study is the lack of any test of statistical significance. Inclusion of such a test is possible, but would require complex methodology that would substantially increase the burden to readers, and might reduce the level of interest in this important topic. Statistical optimization of feature extraction and numerical scores is a non-trivial analytic challenge that deserves greater attention in publication. It was thought that limiting this project to a correlation study, and the presentation of high dimensional analytic results in the form of a three graphic plots, might serve to maintain the readability and clarity among interested readers. Data from this analysis support the validity of the BIBR as a reasonable solution, and suggest that no minimum constraint ratio is needed for automated analysis methods. (Computer scoring algorithms have already presumably made use of any measurable difference that could be extracted from relevant and comparison questions.) Another limitation of this analysis is the lack of a second sample with which to compare these results. Continued interest and research is recommended in the optimization of feature extraction and numerical transformation for both automated and manual test data analysis methods.

References

Kircher, J.C., Horowitz, S.W, and Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior, 12*(1), 79 – 90.

Krapohl, D. & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph, 28*, 209-222.

Krapohl, D.J. (2002). Short report: Update for the Objective Scoring System. *Polygraph, 31(*4), 298 – 302.

National Center for Credibility Assessment (2017). *Test Data Analysis: Numerical Evaluation Scoring System Pamphlet*. Available from the author. (Retrieved from http://www.antipolygraph.org on 4-13-2019).

Nelson, R. & Krapohl, D. K. (2017) Practical polygraph: a recommendation for combining the upper and lower respiration data for a single respiration score. *APA Magazine. 50*(6), 31-41.

Nelson, R., Krapohl, D.J., and Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph, 37*(3), 185-215.

Nelson, R. & Handler, M. (2015). Statistical reference distributions for comparison question polygraphs. *Polygraph, 44*(1), 91-114.

Nelson, R. & Handler, M. (2018). Practical polygraph: seven things to know about feature extraction with electrodermal and cardio data. *APA Magazine, 51*(1), 116-132.

Nelson, R. (2017). Multinomial reference distributions for the Empirical Scoring System. *Polygraph and Forensic Credibility Assessment, 46* (2), 81-115.

Nelson, R. (2018). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. *Polygraph and Forensic Credibility Assessment, 47*(2), 158-175.

Peng, R. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Using Virtual Reality to Improve Memory Recall and Detection of Deception in Forensic Interviews

Joyce Yan Ting Sam[1]

Lin Qiu[1]

and Ky Phong Mai[1]

Division of Psychology, School of Social Sciences, Nanyang Technological University, Singapore

Abstract

Virtual Reality (VR) bears promising contributions to the polygraph community given its ability to reconstruct virtual environments modeled after real-world settings during forensic interviews. It provides the possibility to transport interviewees (e.g., eyewitnesses and suspects) into virtual crime scenes, thus benefiting their recall of the crime event by reinstating the original environment. The occurrence of contaminated false confessions may also be minimized as "inside information" that are only known to investigators and true culprits can be obviated from these virtual crime scenes. Apart from memory enhancement, detection of deception may also take place within these virtual forensic interviews with the help of eye-tracking technology. In this paper, we provide empirical evidence from two preliminary studies suggesting that virtual forensic interviews may produce similar spatial memory performance with those obtained from crime scene visitations. Moreover, our results suggest that forensic interviews in a virtual crime scene may lead to better spatial recall as compared to the use of crime scene photographs in an interview room. We conclude by providing insights on how deception can be detected in virtual interviews. By enhancing interviewees' memory and detection of deception, VR presents great potential for improving the accuracy and reliability of information gathered in forensic interviews.

Keywords: *Virtual Reality, forensic interviews, eyewitness memory, contaminated false confessions, detection of deception*

Use of Virtual Reality to Improve Memory Recall and Detection of Deception

Over the past decade, Virtual Reality (VR) has garnered growing attention from practitioners and researchers alike. For example, a non-exhaustive list of VR applications includes games and entertainment (Zyda, 2005), social skills training (Didehbani et al., 2016), military training (Alexander et al., 2017), surgical simulations (Gallagher et al., 2005), treatment of mental illnesses (Freeman et al., 2017), as well as education and learning (Vesisenaho et al., 2019). Additionally, VR has also been

[1] Joyce Y. T. Sam https://orcid.org/0000-0003-0795-8601

Lin Qiu https://orcid.org/0000-0002-3587-5371

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Correspondence concerning this article should be addressed to Joyce Yan Ting Sam. Email: SAMY0004@e.ntu.edu.sg

increasingly adopted as a research methodology due to its capability to create ecologically valid, secure, and realistic environments (Morganti, 2004).

Using computer technology, VR transports people to immersive virtual environments (IVEs) where exploration and interaction with the immediate surrounding can freely occur (Bekele & Champion, 2019; Herrera et al., 2018). By replacing perceptual input from the reality with those from the virtual world, the physical world is completely blocked out (Bailenson, 2018). This creates in users the feeling that they are inside the virtual world. Given VR's ability to recreate experiences with a high level of realism (Cipresso et al., 2018), crime scene simulations and pre-test forensic interviews in VR are made possible. Although the current practice for conducting forensic interviews typically revolves around the use of crime scene photographs and field interviews (Forensic Science Bureau, 2019; Geberth, 2015; Gehl & Plecas, 2017), interviewees (e.g., eyewitnesses, suspects, etc.) may be transported into virtual crime scenes to provide their accounts in the future. Notably, this means that witnesses can be immersed in highly realistic 3D environments that can be controlled. Two potential benefits are likely to follow from the employment of virtual forensic interviews – the enhancement of interviewees' memory and detecting deception.

In the remaining of this paper, we first present an introduction to human memory and a review of the theoretical basis underpinning the effective use of VR for improving memory. Then, we provide results from two preliminary studies that demonstrate VR's potential for memory enhancement. Finally, we shed light on how eye tracking in VR can improve detection of deception by reviewing previous work in this area.

Utilizing VR for Memory Enhancement

The Human Memory

Early research on memory has long revealed that human memory is not a single entity but comprises of multiple types (Squire, 2004; Squire & Dede, 2015). An important distinction is that of short-term memory (STM) and long-term memory (LTM) (Atkinson & Shiffrin, 1968; Atkinson & Shiffrin, 1971), with the former storing a small amount of information for a very brief period of time (i.e., up to 18 seconds) and the latter storing an unlimited amount of information for an indefinite duration. The term "short-term memory" is often used interchangeably with the term "working memory" despite not being entirely the same (Aben et al., 2012; Norris, 2017). Although researchers hold dissenting views as to what working memory constitutes, a common agreement is that STM is simply a temporary information store whereas working memory is an active system that temporarily stores and manipulates information (see Aben et al., 2012 for a review on STM and working memory). For instance, Baddeley (1992) regards working memory as the maintenance and manipulation of necessary information for tasks that are cognitively complex (e.g., learning and language comprehension). In one of the most well-known experimental neurosurgeries in history, patient H.M. exhibited an inability to recall previously seen faces, scenes, and words after a medial temporal lobe bilateral resection to control seizures (Scoville & Milner, 1957; Squire, 2009). Albeit so, Milner (1962) discovered that he was able to master a task (i.e., mirror drawing) which required hand-eye coordination without any recollection of having practiced it before (as cited in Squire & Dede, 2015). This unexpected finding suggests that there are different forms of long-term memory.

Indeed, the fact that long-term memory can be categorized into two major memory systems – nondeclarative (i.e., implicit) and declarative (i.e., explicit) memory – has been supported by numerous studies (e.g., Cohen & Squire, 1980; Packard et al., 1989; Schacter & Buckner, 1998). Nondeclarative memory is an umbrella term used to denote multiple forms of memory that are inaccessible to conscious awareness (Kandel & Squire, 1999; Squire, 1987; Squire & Dede, 2015; Squire & Zola-Morgan, 1988), such as motor and perceptual skills (e.g., driving a car), priming and perceptual learning (e.g., detection of visual stimuli), simple classical conditioning (e.g., fear of aversive stimuli), as well as nonassociative learning (e.g., sensitization and habituation) (see Squire & Dede, 2015 for a review on nondeclarative memory). Nondeclarative memory manifests

itself through performance and what has been learnt is shaped by past experiences without requiring any conscious remembrance or memory content (Squire, 2004; Squire, 2009; Squire & Dede, 2015). Neuroanatomically, nondeclarative memory relies predominantly on the amygdala, striatum, cerebellum, and neocortex (Kandel et al., 2014; Squire & Dede, 2015).

Declarative memory, on the contrary, represents memory for events (i.e., episodic) and facts (i.e., semantic) (Kandel et al., 2014; Kandel & Squire, 1999; Squire, 2004; Squire, 2009; Squire & Dede, 2015; Tulving, 1983). Unlike nondeclarative memory, it is expressed through recollection and serves as a way of modeling knowledge about the external world. As the material remembered are accessible to conscious awareness and can be juxtaposed, declarative memory can guide performance across multiple different test conditions and contexts (Squire, 2004; Squire & Dede, 2015). The key brain regions involved in declarative memory are the hippocampus and the parahippocampal gyrus consisting of the entorhinal, perirhinal, and parahippocampal cortices (Squire, 1992; Squire & Zola-Morgan, 1991). Relatedly, attention has been found to be a critical factor for encoding and retrieving of declarative information due to its contribution to the stabilization of hippocampal representations (Aly & Turk-Browne, 2016; Muzzio et al., 2009). Following the above definitions, it appears that memory for a crime event is a kind of long-term episodic memory as witnesses and suspects are required to recount the time-and place-specific incident (Gavin, 2014).

Sources of Memory Failures

Eyewitness memories provide direct evidence of how a crime unfolds (Albright, 2017; Wells et al., 2006). Yet, eyewitness testimonies often fail egregiously, leading to mistaken identifications and innocent convictions (Garrett, 2011; Innocence Project, n.d.). Similarly, not all confessions from suspects are accurate and true (Inbau et al., 2013). According to the Innocence Project (n.d.), false confessions are involved in approximately 30% of all wrongful conviction cases exonerated by DNA testing (as cited in Kassin, 2014). Clearly, memory is fallible and susceptible to errors.

Schacter (1999; 2001) posited that memory failures can be categorized into seven distinct fundamental "sins": transience (decreasing accessibility of memory with time), absent-mindedness (forgetting due to inadequate attention during encoding or retrieval), blocking (temporary inaccessibility of encoded information), misattribution (false assignment of memory to incorrect sources), suggestibility (integration of false information into memory due to external influences), bias (distorted memories of past events due to present beliefs and knowledge), and persistence (pervasive recollections of unwanted information such as traumatic experiences). The first three transgressions are regarded as "sins of omission" that entail forgetting, whereas the last four transgressions are deemed as "sins of commission" that involve unwanted or inaccurate memories (Murray, 2003). Of particular importance to the crime context are the sins of misattribution and suggestibility.

Firstly, misattribution occurs when individuals falsely attributes recollections to an inaccurate source (Murray, 2003; Schacter, 1999; Schacter, 2001). For example, people may erroneously recall seeing a face in a specific context when it was actually encountered in another (Read, 1994). Such source confusions possess critical implications for eyewitness testimony as a witness may falsely identify a familiar but innocent person as the perpetrator of a crime (Perfect & Harris, 2003). Termed as "unconscious transference", this phenomenon has been demonstrated in multiple studies (e.g., Earles et al., 2008; Loftus, 1976; Ross et al., 1994). To illustrate, participants in Loftus (1976)'s study first heard a narrative describing a transgression. The introduction of each character in the story was accompanied by a photograph. Three days later, participants were instructed to identify the culprit from five photographs. Results showed that an innocent incidental character was more likely to be selected when the real criminal was not included in the lineup identification, thus suggesting that transference errors may happen due to the incorrect attribution of familiarity to a wrong contextual source.

Secondly, suggestibility happens when inaccurate details are incorporated into memory due

to suggestions, comments, or leading questions (Murray, 2003; Schacter, 1999; Schacter, 2001). Similar to misattribution, suggestibility imposes potentially damaging effects on the legal system. One profound consequence concerns the creation of false memories, which in turn may lead to false confessions (Schacter, 1999). An empirical study conducted by Kassin and Kiechel (1996) demonstrates how suggestions at the time of retrieval can produce recollections of events that did not take place. Participants were assigned to either the fast-pace or slow-pace reaction task, which required them to either type a list of letters quickly or slowly. They were explicitly told to not press the "ALT" key as it would cause the computer program to malfunction. Essentially, the program was configured to crash after a minute regardless of whether the key was pressed. Although innocent, participants were accused of having pressed the "ALT" key. Upon denying the allegation, one group heard a confederate witness affirmed to having seen them pressed the key while the other group did not. Surprisingly, almost 70% of all participants falsely admitted to the act despite being innocent. This result was especially pronounced in the fast-pace/witness group, with a 100% false confession rate and 35% of participants confabulating false details of how they committed the error.

Influencing the decision of criminal justice officials to a large extent, false confessions are one of the most incriminating and compelling evidence of guilt against an innocent defender (Leo, 2009). These false confessions often come across as believable and credible as it is difficult to fathom why an innocent person would confess. According to Garrett (2015), many of the false confessions revealed through DNA testing were particularly detailed and included "inside information" about the crime that only true culprits would know. In fact, 24 out of the 26 exoneration cases comprising false confession in the past decade involved crime scene information that were consistent with crime scene details which only the perpetrator and investigators could have known. Clearly, the false confessions of these innocent convicts were contaminated by details leaked during the interrogative interviews. Police investigators may aid in the creation of false confessions by intentionally or unintentionally suggesting facts and details of the crime and/or crime scene to these suspects, thus contaminating their postadmission narrative (Garrett, 2015; Leo, 2009). In a similar fashion, bringing suspects to the crime scene as a part of the investigation may also expose them to these "inside information" which could later be falsely incorporated into their memory and confessions.

Context-dependent Memory

Apart from its susceptibility to errors, memory is also a cognitive process sensitive to changes in context (Robin & Moscovitch, 2013; Smith & Vela, 2001). Specifically, context-dependent memory pertains to the phenomena where matching of contexts at encoding and retrieval results in an enhanced recall of specific information (Grant et al., 1998). According to Smith (2007), "context, most generally defined, is that which surrounds" (p. 111). This vague definition suggests that a "context" can refer to anything that encompasses a focal stimulus, such as environmental settings and internal states (Løhre, 2011).

Notably, the effect of environmental context on memory performance has been examined extensively (e.g., Smith, 1986; Smith et al., 1978). In a landmark study of environmental context-dependent memory, Godden and Baddeley (1975) found that divers who memorized word lists underwater or on land exhibited better recall memory when they were tested in the original learning environment. Contrarily, testing in a novel setting produced a weaker recall of the learnt words. This finding demonstrates that memory recall can be improved by a match in the environmental contexts during encoding and retrieval. Similar results were reported in studies using room manipulations (McDaniel et al., 1989; Smith, 1979), odors (Cann & Ross, 1989), and music (Balch et al., 1992; Smith, 1985).

One pivotal practical implication of environmental context-dependent memory is eyewitness testimony. In their study, Smith and Vela (1992) instructed participants to watch a mock crime before completing a lineup identification of the perpetrator either in the same room or a different room. Recognition accuracy was found to be higher when the lineup identifica-

tion was administered in the same room than in the different room, therefore providing empirical evidence that eyewitness memory can be enhanced by reinstating the crime context and returning witnesses to the crime scene. In fact, encouraging mental reinstatement of the environment during crime has been established as a key rule for conducting cognitive interviews to elicit information from eyewitnesses (Geiselman et al., 1985), as well as a component in the PEACE Model of Investigative Interviewing (Association of Chief Police Officers in England and Wales [ACPO], 2001). Based on Tulving's (1979) specificity encoding principle, the most amount of relevant information will be remembered when there is a maximal overlap between the context in which the crime occurred and the context in which recall was made.

Virtual Crime Scenes

Instinctively, returning interviewees (i.e., eyewitnesses and suspects) to the crime scene appears to be the best approach for environmental reinstatement and improving recollection of the transgression. It is essential, however, to recognize that this may also incur in them false memories of the crime event. As previously mentioned, crime scene visitations may disclose "inside information" concerning the crime site, thereby contaminating the false confessions of innocent suspects. Considering both the memory-enhancing effect of context reinstatement and how the crime scene may be a source of leakage of "inside information", interviewees' memory of the crime can be best ameliorated by returning to a crime scene that does not contain any critical details. With the ability to control and manipulate what a virtual environment will display, VR can be utilized for recreating the crime scene whilst obviating all "inside information" that should be held back from the public. Interviewees can therefore be provided with adequate contextual cues for memory retrieval while being obscure to key details about the crime scene.

Aside from the aforementioned, virtual crime scenes can be useful when visitations to the original crime scene is impossible. To illustrate, the crime scene may have changed significantly since the crime took place (e.g., due to construction) (Bailenson et al., 2006) or destroyed in the case of an arson or bad weather (e.g., ongoing heavy rain) (Fish et al., 2011). Another advantage of using virtual crime scenes concerns its capability to provide a safe environment for subjects to relive the crime (Dath, 2017). Providing an invulnerable environment to be in can be particularly important when witnesses are hesitant and fearful of returning to the actual crime site due to the possible trauma from witnessing a crime (e.g., murder).

A critical concept worth discussing is the method of developing these virtual crime scenes. Specifically, they can be accurately recreated through mobile laser scanners and photogrammetry techniques (see Dath, 2017; Sieberth et al., 2019 for a review). Additionally, inputs from crime scene photographs or Google Earth can also provide crucial environmental information for the reconstruction of virtual crime sites. With respect to the latter, Google Earth allows images of landscapes and locations all around the globe to be captured from various angles through satellite imagery (Yu & Gong, 2012). For crimes that happened in public places, virtual crime scenes can thus be created by modeling after the corresponding satellite images from Google Earth.

The Present Research

Given the vital role of contextual environment in accurate recollection of a crime event, two preliminary studies were conducted to examine how memory of previously encoded information in the real world plays out in a virtual replication of the same environment. Specifically, a memory task in which participants had to recall the locations of various objects in a room was employed. Both studies adopted a similar experimental design and methodology, albeit using different stimuli and environmental settings.

Pilot Study 1

Pilot Study 1 was first carried out to determine if there are preliminary support for utilizing VR to enhance memory. Since IVEs designed after

real-life settings should presumably provide sufficient contextual information for memory recall, we hypothesized that individuals will be equally able to remember previously seen objects in a real environment when being tested in the actual environment and its virtual replica. In accordance with previous research on environmental context-dependent memory (e.g., Godden & Baddeley, 1975; Smith, 1979), it was also predicted that recall in a novel location (i.e., different from the encoding environment) will result in the worst memory performance.

Hypothesis 1: Recall performance in the virtual replica of the encoding environment is equivalent to that in the actual encoding environment.

Hypothesis 2: Recall performance in the novel environment is worse than that in the actual encoding environment and its virtual replication.

Two object recognition tasks and one object location task were administered to obtain a comprehensive measure of memory performance. The former predominantly assess visual memory whereas the latter mainly measures spatial memory (Gamberini, 2000). Particularly, the second object recognition task assesses the ability to recognize item-specific information.

According to the fuzzy-trace theory (FTT; Reyna & Brainerd, 1995), item-specific information refers to the characteristics and details of an item (e.g., a specific mug or book) (Andermane & Bowers, 2015; Hudon et al., 2006).

Method

All materials and procedures were approved by the ethics committee of the Psychology Program in Nanyang Technological University (NTU). Informed consent was obtained from all participants.

Participants. Six participants (6 females, 0 males; M_{age} = 22.83, SD_{age} = 1.72) were recruited from NTU. They were told that the study aimed to explore human attention and cognition in VR. Participants were aware that they might be required to experience VR individually. Following their completion of two 30-minutes experimental sessions, they were rewarded with a remuneration of $10. Of these participants, four were Chinese (66.7%), one was Malay (16.7%), and one was Arabian (16.7%). All participants did not experience motion sickness, vertigo, or seizures and possess either a normal vision or corrected-to-normal vision.

Design and Procedure. Adopting a between-subject design, participants were randomly assigned to either of three retrieval conditions: same room (n = 2), VR room (n = 2), or photograph (n = 2). The experiment consisted of two sessions – an encoding session and a retrieval session. In the encoding session, participants were first brought into a room where they reported their age, gender and ethnicity. Ten objects commonly found in an office setting, such as a textbook, mug, and poster, were placed at different locations within the room (see Figure 1). These objects served as the target objects in this study. Items that are inconsistent with an office schema were not employed as expectation has been found to influence memory (e.g., Friedman, 1979; Maki, 1990). Compared to expected information in a scene, people may remember unexpected information better as these details tend to garner more attention during encoding. To ensure that participants saw and attended to the various target objects, an attention task was administered. In particular, participants were provided with a list containing the target objects and were instructed to check if a sticker is present on each object (see Appendix A). The two posters were excluded from the attention task as we did not intend for participants to remove them from the wall. A yes/no response format was used to indicate the presence of the sticker. A sticker was placed on the following target objects: mug, textbook, potted plant, water bottle, and waste bin. Upon completion of the attention task, participants were thanked and reminded to return for a retrieval session on the next day.

Figure 1 The Encoding Room and Locations of Target Objects

The retrieval session was carried out 24 hours later. Depending on the condition they were randomly assigned to, participants reported to different venues. In the same room condition, participants entered the encoding room where number labels have replaced the target objects seen in the encoding session (see Figure 2). In the VR room condition, participants wore an HMD (HTC Vive Pro) and were transported into a virtual replica of the encoding room created using the Unity software (version 2018.2.18f1) (see Figure 3). Likewise, participants saw number labels instead of the target objects. In the photograph condition, participants were brought into a novel room (see Figure 4) and given a photograph of the encoding room to refer to (see Figure 5). The experimental procedure across all three conditions was similar. All participants completed three different memory tasks – a free-choice object recognition test, a forced-choice object recognition test, and an object location test.

Figure 2 Retrieval Room for Same Room Condition

Figure 3 Retrieval Room for VR Room Condition

Figure 4 Retrieval Room for Photograph Condition

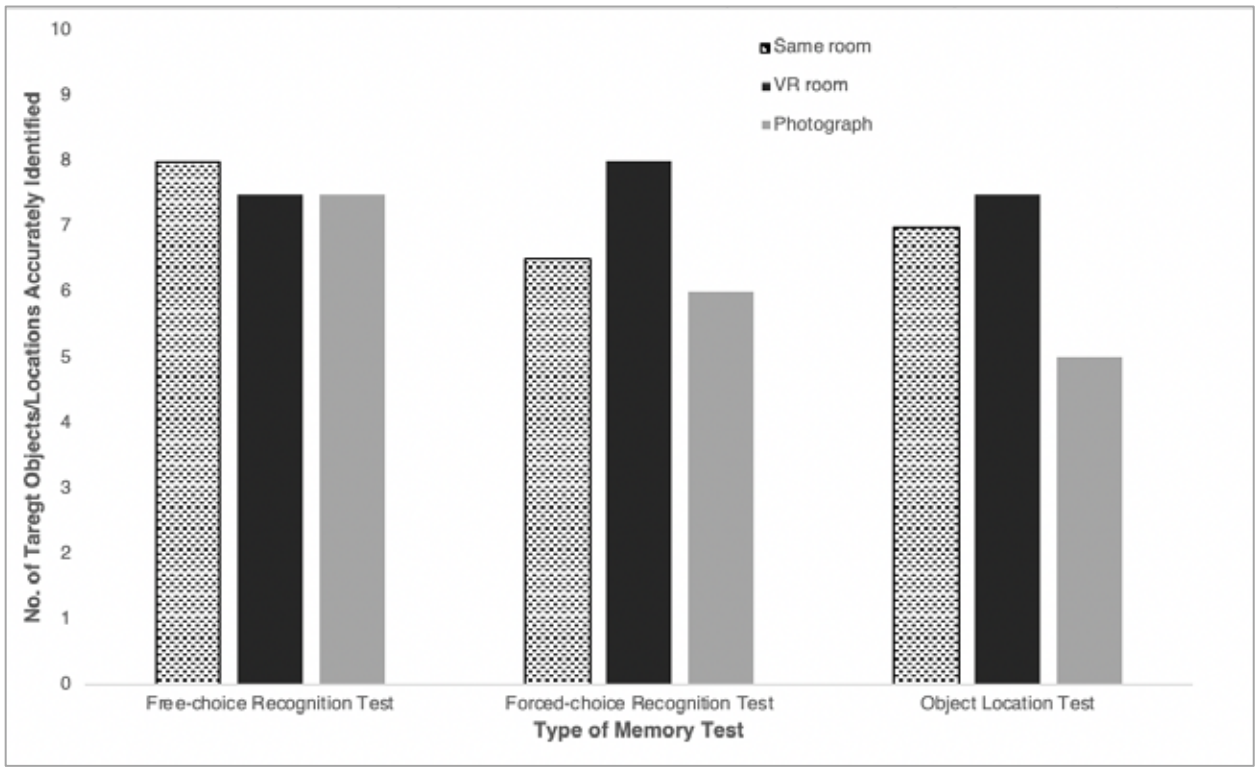
Figure 5 A Photo of the Encoding Room Presented to Participants in the Photograph Condition

In the free-choice recognition test, participants were given a 16-item checklist which contains the 10 target objects alongside 6 new objects that were absent during the encoding session (see Appendix B). Participants were tasked to identify the objects which they remembered seeing during the encoding session. In the forced-choice recognition test, participants were given four possible options (e.g., four mugs with different designs) for each target object (see Appendix C) and were instructed to identify the object they remembered seeing during the encoding session. In the object location test, participants were presented with an image of each target object and were told to identify its location (see Appendix D). The same room and VR room conditions referred to the number labels in their immediate surroundings, while the photograph condition referred to the photograph of the encoding room. Participants verbalized their responses on all three memory tasks, which were recorded and coded for accuracy. Upon completion of the memory tasks, participants were debriefed, thanked, and remunerated for their participation.

Results and Discussion

Examining the mean memory performance across the three conditions, the same room condition (M = 8.00) accurately identified more target objects in the free-choice recognition test compared to the VR room and photograph conditions (VR room: M = 7.50; photograph: M = 7.50). In the forced-choice recognition test, the VR room condition (M = 8.00) accurately identified more target objects as compared to the same room and photograph conditions (same room: M = 6.50; photograph: M = 6.00). In the object location test, the VR room condition (M = 7.50) accurately identified more target object locations as compared to the same room and photograph conditions (same room: M = 7.00; photograph: M = 5.00). Figure 6 illustrates the results.

Figure 6 Mean Memory Performance on All Three Memory Tests Across Conditions



Among the three memory tasks, results from the object location test provides the strongest preliminary support for our hypotheses. Similar memory performance was found across the same room and VR room conditions, albeit the latter showing a slight improvement. In addition, memory recall in the photograph condition was weaker as compared to both the same room and VR room conditions. Since object recognition tests typically assess visual memory whereas object location tests typically measure spatial memory, our findings suggest that spatial memory performance was more consistent with our predictions as compared to visual memory performance.

Study 2

Although promising results were obtained from Pilot Study 1, the inclusion of 10 target objects might have imposed a low demand on participants' memory. Consequentially, they might have found the memory tasks too easy. Study 2, therefore, aims to increase the difficulty of the memory tasks by conducting the experiment in a larger room and by increasing the number of target objects.

Method

Participants. Twenty-one participants (10 females; M_{age} = 25.48, SD_{age} = 3.09) were recruited from NTU. Likewise, they were told that the purpose of the study was to investigate attention and cognition in VR. Participants were tested individually and were rewarded with a $10 remuneration following their completion of the encoding and retrieval sessions. Of these participants, eight were Chinese (38.1%), ten were Indian (47.6%), one was Eurasian (4.8%), one was Filipino (4.8%), and one was Russian (4.8%). No participants experienced motion sickness, vertigo, or seizures and all of them possess either a normal vision or corrected-to-normal vision.

Design and Procedure. Adopting the same between-subject design, participants were randomly assigned to either the same room (n = 7), VR room (n = 7), or photograph (n =

7) condition. All procedures were identical to those in Pilot Study 1 (see Appendix E, F, G, H for attention task and memory tasks used). However, the experiment was conducted in a larger room (see Figure 7) and the number of target objects was increased from 10 to 22 (see Figure 8). The 22 target objects were placed at 15 different locations such that there could be multiple objects at one location (e.g., duct tape and pins are placed at location 3).

Figure 7 Encoding Room and the Three Retrieval Conditions

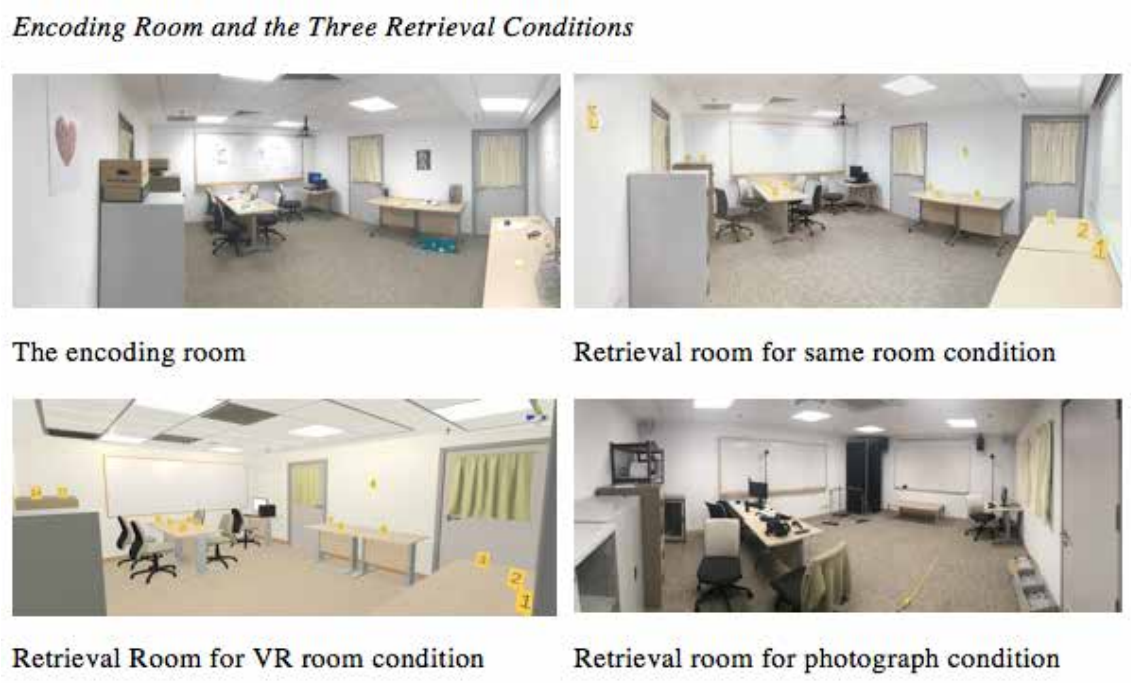
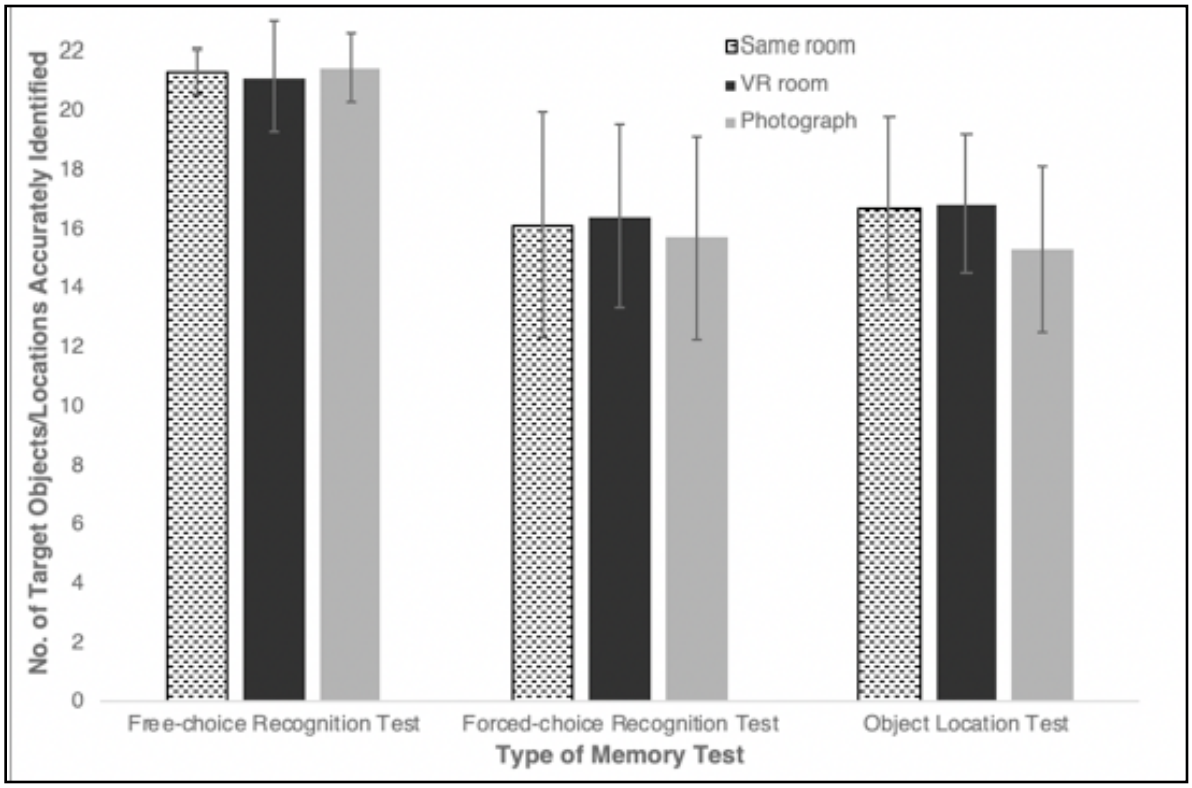


Figure 8 Location of Target Objects

Results and Discussion

As in Pilot Study 1, we examined the mean number of target objects/locations successfully recalled in each condition. In the free-choice recognition test, all three conditions accurately recalled a similar number of target objects (same room: *M* = 21.29, *SD* = 0.76; VR room: *M* = 21.14, *SD* = 1.86; photograph: *M* = 21.43, *SD* = 1.13). Likewise, all three conditions accurately remembered a similar number of target objects in the forced-choice recognition test (same room: *M* = 16.14, *SD* = 3.81; VR room condition: *M* = 16.43, *SD* = 3.10; photograph: *M* = 15.71, *SD* = 3.40), albeit with the photograph condition performing slightly worse. For the object location test, we adopted a clustered location analysis as the target objects were placed in close proximity to each other and examining the memory for exact target object locations would have led to misleading findings. Namely, the failure to recall specific location of the target objects may not necessarily indicate poor memory. Participants might have remembered the objects' general (i.e., on the cabinet) but not its precise location (i.e., on the top left corner of the cabinet). Apart from the two posters that were on the walls, we clustered together the remaining locations that were near to each other. Four clusters emerged as a result: cluster 1 comprised of locations 1, 2 and 3, cluster 2 comprised of locations 4, 6, and 7, cluster 3 comprised of locations 8, 9, 10, 11, and 12, and cluster 4 comprised of locations 13 and 14. Analyzing memory performance based on these clustered locations, the same room (*M* = 16.71, *SD* = 3.09) and VR room conditions (*M* = 16.86, *SD* = 2.34) accurately identified a similar number of target object locations. The photograph condition, on the other hand, accurately identified fewer target object locations (*M* = 15.29, *SD* = 2.81). Figure 9 illustrates the results.

Figure 9 Mean Memory Performance on All Three Memory Tests Across Conditions



Once more, results on the object location test was the most in line with our hypotheses. The overall trend is still the same, with similar spatial memory performance observed across both the same room and VR room conditions, as well as a weaker spatial recall exhibited in the photograph condition. This pattern of results is consistent with our findings from Pilot Study 1.

General Discussion

Across two preliminary studies, we examined memory performance in real-world and virtual environments. In particular, memory performance on three different tasks was compared across three different retrieval environments (i.e., same room, VR room, and novel room). In both studies, performance on the object location test provided the strongest support for the prediction that memory recall in a virtual replication of the encoding environment will be similar to that in the actual encoding environment. Moreover, results on the object location test also confirmed our expectation that memory retrieval in a novel setting will produce the worst memory as compared to the actual encoding environment and its VR replica. Particularly in the case of spatial memory, the collective findings from Pilot Study 1 and Study 2 suggest that virtual crime scenes may be equally effective as crime scene visitations for producing accurate recall of the crime event. Furthermore, these findings also propound that VR may improve interviewees' memory as compared to the use of photographs, which are commonly employed in investigative interviews for gathering information regarding the crime. By helping interviewees to better recall the crime event, virtual forensic interviews may thus lead to testimonies of greater accuracy and reliability.

One explanation for the pattern of results observed may be the spatial relativity offered in VR (Briggs, 2018; Krokos et al., 2018). To illustrate, the ability to integrate one's own vestibular and proprioceptive inputs (e.g., overall sense of body movement, acceleration, and position) may serve as an additional retrieval cue during the memory process. Simply put, individuals are able to use their own physical presence as a reference point for retrieving the location of different objects. In the context of our studies, participants in the same room and VR room conditions were able to use their body position as a frame of reference in recalling the location of target objects previously seen. Conversely, instructing participants to complete the object location test with reference to a photograph (i.e., photograph condition) denied them of locomotion and the capability to utilize their physical presence as a retrieval cue. The use of VR to enhance spatial relativity is consistent with the change perspective component of the cognitive interview technique where eyewitnesses and victims of crime are encouraged to recount the incident from a different perspective (e.g., those of other witnesses or someone else present at the scene) (Geiselman et al., 1984). Through VR, interviewees are able to adopt another person's frame of reference in recalling the key crime facts.

There are a number of limitations to our studies. For one, both studies adopted small sample sizes. This is because our pilot studies are works-in-progress and were conducted to determine if there are preliminary support for our predictions. Based on meta-analytic results indicating an effect size of d = 0.23 between reinstatement effect and environmental context-dependent memory (Smith & Vela, 2001), a statistical power analysis performed with G*Power 3.1 (Faul et al., 2009) for alpha = .05 and power = 0.80 indicated a sample size estimate of N = 186 for a three-group comparison. Future studies will thus recruit at least 186 participants to meet the required sample size. In addition, we recognize that no male participants were recruited in Pilot Study 1 despite past literature documenting various sex differences in memory (e.g., Davis, 1999; Herlitz & Rehnman, 2008; Skowronski et al., 1991). Our results from Study 2 were in fact, consistent with previous research (e.g., Herlitz & Rehnman, 2008) showing that women are better at remembering objects whereas men are better at visuospatial processing. Female participants in Study 2 outperformed male participants on the forced-choice recognition test (female: M = 17.30, SD = 2.83; male: M = 15.09; SD = 3.53) while male participants outperformed female participants on the object location test (female: M = 15.00, SD = 2.98; male: M = 17.45; SD = 1.92). These findings suggest that sex differences in memory should not be discounted and future studies should recruit participants of both sexes.

Secondly, it is possible that our attention task might have heightened participants' awareness about the study's true purpose. Solely instructing participants to search for stickers on the target objects during the encoding session might have produced suspicions about the study's real aims. Subsequently, this might have led to an intentional encoding of the target objects which would have affected the data collected. Since recollection of crime events are often incidental (Kausler, 1991), this would reduce the generalizability of the study's results to real-world applications. As we did not examine if participants could guess the true intent of our study, future work should consider doing so to ensure that all participants were unaware of the study's real objective.

Another limitation concerns the lack of free recall measure in our studies. Typical investigative interviews involve the use of free narratives at the beginning as free recall has been shown to provide the most accurate recollection of the crime (Lipton, 1977; Snook et al., 2012). Moreover, these free narratives allow investigators to gather insights on the interviewees' mental representation of the crime, which in turn aids them in structuring the successive questions of the interview (Milne et al., 2007). Without examining free recall performance, our studies were unable to shed light on how this important form of memory plays out in VR. Hence, future studies can fill this research gap by exploring free recall performance across real-world and virtual settings.

Lastly, our studies are different to real crime events as participants remembered neutral, everyday objects in a room, which are likely to have induced a neutral mood and low arousal levels. however, actual crimes are often emotionally arousing as witnesses encode information under stressful circumstances (Hoscheidt et al., 2014). With past research documenting the various ways in which emotional arousal and mood may affect eyewitness memory (e.g., Houston et al., 2013; Thorley et al., 2016), memory performance in VR may differ for those of crime events and neutral object. Thus, future work should investigate memory for a crime event witnessed in real world. One particular approach to do so is to present participants with a mock crime (e.g., theft) and test their memory for the event in the actual crime scene and its virtual counterpart. This will provide information as to whether placing eyewitnesses into a virtual crime scene can improve their memory for a crime witnessed in reality.

Utilizing VR for Detecting Deception

Empirical research hinting at the possibility of detecting deception in VR has recently emerged. For instance, Mapala and team (2017) randomly assigned participants into a deception or non-deception condition. In the former, participants were asked to place into a backpack prohibited items that are not allowed onto planes, such as a hammer and a pair of scissors. In the latter, however, participants were told to put into a backpack non-prohibited items, such as a chocolate bar and a t-shirt. Thereafter, all participants were transported into a virtual airport security checkpoint where they had to answer yes/no questions prompted by an avatar security guard. Questions included those that were truthful (e.g., "Are we currently in an airport?") and deceptive (e.g., "Are you sure there are no restricted items in your bag?") in nature. In line with previous deception literature (Kircher & Raskin, 2016), results revealed that liars responded quicker when answering deceptive questions while the response latencies of truth tellers were similar when answering both truthful and deceptive questions. According to the researchers, the deceivers' shorter response latencies might have been a conscious or subconscious attempt to appear more confident in front of the security guard. This study puts forth an example of how pre-test forensic interviews can be conducted within an IVE to detect deception.

Measuring eye movements and pupillary responses to detect regions of interest within a person's field of view (Gonzalez-Sanchez et al., 2017), eye tracking technology shows apparent potential for use with VR to detect deception. As a matter of fact, the eyes have been long known to offer at least three different deceptive cues –fixation points (Derrick et al., 2010), blinks (Fukuda, 2001), and pupil dilation (Dioniso et al., 2001). Firstly, fixations points represent one of the most prominent eye gaze pattern measure that has been empirically examined in detection of deception studies (Schuetzler, 2012). To illustrate, par-

ticipants in Derrick et al.'s (2010) study were either asked to construct a counterfeit bomb or received no such instructions. Whilst tracking their gaze patterns, both groups of participants later underwent a guilty knowledge test and were presented with altered pictures of explosive devices which they have or have not built. A subsequent analysis of fixation points revealed that guilty participants (i.e., those who constructed the "bomb") gazed much more at the modified portion of the image as compared to innocent participants (i.e., those who did not construct the "bomb"), hence demonstrating the effectiveness of using fixation points for distinguishing deceptive participants from truthful ones.

Secondly, eyeblink measures (e.g., blink rate and blink latency) have also received considerable attention in the deception literature given its ability to reflect cognitive processes and effort (Goldstein et al., 1985; Schuetzler, 2012; Stern et al., 1984). Using a guilty knowledge card test, Fukuda (2001) tasked participants to select one playing card from a set of five. The chosen card serves as the relevant stimulus whereas the unselected ones represent irrelevant stimuli. After which, participants were presented with all five cards and were explicitly instructed to lie about the card that they chose. Eye blink recordings showed that blink rates were significantly lower when participants viewed the relevant card in comparison to the irrelevant ones. Blink latency, which refers to the time between the first blink and the presentation of the stimulus, was also found to be longer when participants viewed the relevant stimuli.

A third ocular deceptive indicator, pupil dilation is postulated to change depending on an individual's cognitive effort and processing load (see Beatty, 1982 for a review). According to Furedy and colleagues (1988), deceivers must concomitantly attend to both true and false answers when designing lies (as cited in Schuetzler, 2012). This monitoring process requires further cognitive effort, which in turn increases deceivers' pupil diameter in relation to truth tellers. In their study, Dionisio et al. (2001) prompted participants to answer the same set of questions twice while tracking their pupil responses. Particularly, they responded to the questions in both truthful and deceptive manners. Contrary to telling the truth, participants were found to exhibit a greater degree of pupil dilation when confabulating lies and generating deceptive recall. Similar findings were documented by Webb et al. (2009) in their study. Using a mock-crime paradigm alongside the comparison question test (CQT), the pupil diameter of innocent participants was found to increase more when responding to probable-lie questions as compared to relevant questions. In contrast, the pupil diameter of guilty participants did not reflect differential responding to both types of questions. As innocent subjects are presumed to be truthful when answering relevant questions (e.g., Did you take the missing money?) but deceptive when answering probable-lie comparison questions (e.g., Did you ever take something that did not belong to you before the age of 30?) in the CQT, changes in the pupil diameter of innocent participants demonstrate that deception does indeed require more cognitive effort than telling the truth. Since lying on both probable-lie comparison and relevant questions require a similar amount of cognitive effort, the pupil diameter of guilty participants was thus comparable regardless of the question type. Taken together, these studies converge on the finding that lying is undeniably more cognitively demanding than being truthful (Vrij et al., 2006; Vrij et al., 2011).

Developed based on this assumption, the ocular-motor deception test (ODT) (Cook et al., 2012) represents a perfect illustration of the practical use of eye tracking in detecting deception. As opposed to its traditional counterpart, the ODT is a computerized test that can be completed within 40 minutes (Bovard et al., 2019; Kircher, 2018; Kircher & Raskin, 2016; Patnaik et al., 2016). Examinees first receives auditory and written instructions before being presented with a set of true/false statements regarding their possible participation in the crime. Responses to each of these statements are made by pressing a key on the keyboard. Crucially, examinees' ocular behaviors (e.g., pupil responses and eye movements) are recorded while the incriminating statements are read. These ocular-related data are subsequently processed by the computer and a conclusion of an examinee as truthful or deceptive is reached. More specifically, the ODT utilizes

the Relevant Comparison Test (RCT) to determine if examinees are lying. In this test format, questions concerning two relevant issues are combined with neutral questions and the discrepancy between reactions to these two sets of questions informs if an examinee is telling the truth or lying on either of the relevant topics. An examinee who attempts to deceive on either of these relevant issues is expected to show stronger ocular-motor reactions when responding to the questions regarding the respective issue, whereas an examinee who is truthful to both relevant issues is likely to exhibit no significant differences in reactions when responding to both set of questions.

Empirical evidence for the effectiveness of the ODT in detecting deception has been provided by Cook and her colleagues (2012). Across two experiments, guilty participants who have committed a mock crime displayed greater pupil dilation when answering true/false statements deceptively. In comparison to innocent participants, the guilty participants also fixated lesser on the true/false statements, spent lesser time reading, and re-reading them. Notably, overall classification accuracy rates paralleled those of the polygraph, with an 85% accuracy rate in Experiment 1 and an 86% accuracy rate in Experiment 2. The robust efficiency of ODT for detection of deception has since been replicated across cultures (e.g., Patnaik et al., 2016) and in later studies (e.g., Bovard et al., 2019; Patnaik; 2013; Patnaik, 2015).

The foregoing findings collectively exemplify ways in which deception can be detected through the eyes. Taking into account VR's high realism and capacity to simulate real-world settings, it is plausible that incorporating eye tracking into IVEs may constitute an efficacious way for examining eye behaviors and pupillary responses related to deception. Consequently, this means that interviewees may provide their accounts in virtual crime scenes while their eye movements are being tracked and analyzed. These eye tracking inputs may then be employed to uncover deception, thus informing the integrity of the given statements.

Currently, credibility assessment professionals attempt to determine if a person has lied by relying on verbal (e.g., speech content), nonverbal (e.g., body movements) and/or physiological indicators of deception (Vrij et al., 2000). By offering an additional avenue of confirmation for deception, VR eye tracking may therefore provide convergent validity. Aside from enhancing the instrumental detection of deception, it is likely that virtual crime scenes can also aid in the interpersonal detection of deception. For instance, a guilty suspect may be reminded of the transgression when placed in the virtual crime scene, which in turn may lead to pronounced cues during deceptive attempts that could be more easily recognized by investigators. Despite the apparent prospect of VR eye tracking in detecting deception, there is sparse empirical work in this area and further research are required to shed light on its efficacy. Technology is no more than the practical use of science and the near future is bound to witness a proliferation in VR- and eye tracking-related research within the polygraph context.

Conclusion

Due to its immersion and presence, VR allows interviewees to be transported into virtual environments that are recreated after real-life surroundings. Alongside the ability to control the type and amount of details to include in IVEs, "inside information" concerning the crime scene can be omitted in the reconstruction of virtual crime scenes. Consequently, placing eyewitnesses and suspects into these virtual crime scenes may evoke environmental reinstatement and a subsequent improved recall of the crime event while minimizing the likelihood of contaminated false confessions. We provide supporting evidence from two preliminary studies. Besides memory enhancement, detection of deception may also be conducted in VR. Eye tracking may be integrated into VR to distinguish deceptive individuals in virtual environments that are carefully designed. Given its utility, VR and eye tracking represent two frontier technologies that polygraphers can adopt for pre-test information gathering and possible detection of deception.

References

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology, 3*(301), 1–9. https://doi.org/10.3389/fpsyg.2012.00301

Albright, T. D. (2017). Why eyewitnesses fail. *PNAS, 114*(30), 7758–7764. https://doi.org/10.1073/pnas.1706891114

Alexander, T., Westhoven, M., & Conradi, J. (2017). Virtual environments for competency-oriented education and training. In J. I. Kantola, T. Barath, S. Nazir, & T. Andre (Eds.), *Advances in Human Factors, Business Management, Training and Education: Proceedings of the AHFE 2016 International Conference on Human Factors, Business Management and Society, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 23–29). Springer International Publishing.

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *PNAS, 113*(4), 420–429. https://doi.org/10.1073/pnas.1518931113

Andermane, N., & Bowers, J. S. (2015). Detailed and gist-like visual memories are forgotten at similar rates over the course of a week. *Psychonomic Bulletin & Review, 22*(5), 1358–1363. https://doi.org/10.3758/s13423-015-0800-0

Association of Chief Police Officers in England and Wales. (2001). *Investigative interviewing strategy*. National Centre for Policing Excellence.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). Elsevier.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*(2), 82–90. https://doi.org/10.1038/scientificamerican0871-82

Baddeley, A. (1992). Working memory. *Science, 255* (5044), 556–559. https://doi.org/10.1126/science.1736359

Bailenson, J. (2018). Experience on demand: *What virtual reality is, how it works, and what it can do.* W. W. Norton & Company.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Noveck, B. (2006). Courtroom applications of virtual environments, immersive virtual environments, and collaborative virtual environments. *Law & Policy,* 28(2), 249–270. https://doi.org/10.1111/j.1467-9930.2006.00226.x

Balch, W. R., Bowman, K., & Mohler, L. (1992). Music-dependent memory in immediate and delayed word recall. *Memory & Cognition, 20*(1), 21–28. https://doi.org/10.3758/BF03208250

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bekele, M. K., & Champion, E. (2019). A comparison of immersive realities and interaction methods: Cultural learning in virtual heritage. *Frontiers in Robotics and AI, 6(*91), 1–14. https://doi.org/10.3389/frobt.2019.00091

Bovard, P. P., Kircher, J. C., Woltz, D. J., Hacker, D. J., & Cook, A. E. (2019). Effects of direct and indirect questions on the ocular-motor deception test. *Polygraph & Forensic Credibility Assessment, 48*(1), 40–59.

Briggs, S. (2018, June 14). *How virtual reality improves memory.* InformEd. https://www.opencolleges.edu.au/informed/features/virtual-reality-improves-memory/

Cann, A., & Ross, D. A. (1989). Olfactory stimuli as context cues in human memory. *The American Journal of Psychology, 102*(1), 91–102. https://doi.org/10.2307/1423118

Cipresso, P., Giglioli, I. A. C., Raya, M. A., & Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology, 9*(2086), 1–20. https://doi.org/10.3389/fpsyg.2018.02086

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: *Dissociation of knowing how and knowing that. Science, 210*(4466), 207–210. https://doi.org/10.1126/science.7414331

Cook, A. E., Hacker, D. J., Webb, A. K., Osher, D., Kristjansson, S., Woltz, D. J., & Kircher, J. C. (2012). Lyin' eyes: Ocular-motor measures of reading reveal deception. *Journal of Experimental Psychology: Applied, 18*(3), 301–313. https://doi.org/10.1037/a0028307

Dath, C. (2017). *Crime scenes in Virtual Reality – A user centered study* [Unpublished master's thesis]. KTH Royal Institute of Technology Stockholm.

Davis, P. J. (1999). Gender differences in autobiographical memory for childhood emotional experiences. *Journal of Personality and Social Psychology, 76*(3), 498–510. https://doi.org/10.1037/0022-3514.76.3.498

Derrick, D. C., Moffitt, K., & Nunamaker, J. F., Jr. (2010). Eye gaze behavior as a guilty knowledge test: Initial exploration for use in automated, kiosk-based screening. *Proceedings of the Hawaii International Conference on System Sciences.* https://pdfs.semanticscholar.org/87f5/5ce9ff534ea01a460b40ecd78e2edc8d1829.pdf

Didehbani, N., Allen, T., Kandalaft, M., Krawczyk, D., & Chapman, S. (2016). Virtual reality social cognition training for children with high functioning autism. *Computers in Human Behavior, 62*, 703–711. https://doi.org/10.1016/j.chb.2016.04.033

Dionisio, D. P., Granholm, E., Hillix, W. A., & Perrine, W. F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology, 38*(2), 205–211. https://doi.org/10.1111/1469-8986.3820205

Earles, J. L., Kersten, A. W., Curtayne, E. S., & Perle, J. G. (2008). That's the man who did it, or was it a woman? Actor similarity and binding errors in event memory. *Psychonomic Bulletin & Review, 15(*6), 1185–1189. https://doi.org/10.3758/PBR.15.6.1185

Faul, F., Erdfelder, E., Buchner, A., & Lang, A-. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Fish, J. T., Miller, L. S., & Braswell, M. C. (2011). *Crime scene investigation* (2nd ed.). Elsevier.

Forensic Science Bureau. (2019). *Crime scene section: Technical manual.* Austin Police Department. https://www.austintexas.gov/sites/default/files/files/Police/Forensics/CS_Technical_Manual.pdf

Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological Medicine*, 47(14), 2393–2400. https://doi.org/10.1017/S003329171700040X

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108(3), 316–355. https://doi.org/10.1037/0096-3445.108.3.316

Fukuda, K. (2001). Eye blinks: New indices for the detection of deception. *International Journal of Psychophysiology, 40*(3), 239–245. https://doi.org/10.1016/S0167-8760(00)00192-6

Gallagher, A. G., Ritter, E. M., Champion, H., Higgins, G., Fried, M. P., Moses, G., Smith, C. D., & Satava, R. M. (2005). Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. *Annals of Surgery, 24*1(2), 364–372. https://doi.org/10.1097/01.sla.0000151982.85062.80

Gamberini, L. (2000). Virtual Reality as a new research tool for the study of human memory. *Cyberpsychology & Behavior*, 3(3), 337–342. https://doi.org/10.1089/10949310050078779

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong.* Harvard University Press.

Garrett, B. L. (2015). Contaminated confessions revisited. *Virginia Law Review, 101*(2), 395–454.

Gavin, H. (2014). *Criminological and forensic psychology.* SAGE Publications.

Geberth, V. J. (2015). *Practical homicide investigation: Tactics, procedures, and forensic techniques* (5th ed.). CRC Press.

Gehl, R., & Plecas, D. (2017). Introduction to criminal investigation: Processes, practices, and thinking. BCcampus.

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton L. A., Sullivan, S. J., Avetissian, I. V., & Prosk, A. L. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. Journal of Police Science & Administration, 12(1), 74–84.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1985). Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology, 70*(2), 401–412. https://doi.org/10.1037/0021-9010.70.2.401

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goldstein, R., Walrath, L. C., Stern, J. A., & Strock, B. D. (1985). Blink activity in a discrimination task as a function of stimulus modality and schedule of presentation. *Psychophysiology, 22*(6), 629–635. https://doi.org/10.1111/j.1469-8986.1985.tb01658.x

Gonzalez-Sanchez, J., Baydogan, M., Chavez-Echeagaray, M. E., Atkinson, R. K., & Burleson, W. (2017). Affect measurement: A roadmap through approaches, technologies, and data analysis. In M. Jeon (Ed.), *Emotions and affect in human factors and human-computer interaction* (pp. 255–288). Elsevier.

Grant, H. M., Bredahl, L. C., Clay, J., Ferrie, J., Groves, J. E., McDorman, T. A., & Dark, V. J. (1998). Context-dependent memory for meaningful material: Information for students. *Applied Cognitive Psychology, 12*(6), 617–623. https://doi.org/10.1002/(SICI)1099-0720(1998120)12:6<617::AID-ACP542>3.0.CO;2-5

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science, 17*(1), 52–56. https://doi.org/10.1111/j.1467-8721.2008.00547.x

Herrera, F., Bailenson, J., Weisz, E., Ogle, E., & Zaki, J. (2018). Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PLos ONE, 13*(10), 1–37. https://doi.org/10.1371/journal.pone.0204494

Hoscheidt, S. M., LaBar, K. S., Ryan, L., Jacobs, W. J., & Nadel, L. (2014). Encoding negative events under stress: High subjective arousal is related to accurate emotional memory despite misinformation exposure. *Neurobiology of Learning and Memory, 112*, 237–247. https://doi.org/10.1016/j.nlm.2013.09.008

Houston, K. A., Clifford, B. R., Phillips, L. H., & Memon, A. (2013). The emotional eyewitness: The effects of emotion on specific aspects of eyewitness recall and recognition performance. *Emotion, 13*(1), 118–128. https://doi.org/10.1037/a0029220

HTC. (n.d.). *Vive pro eye: Propel your business with precision eye tracking.* Vive. Retrieved May 20, 2020, from https://www.vive.com/eu/product/vive-pro-eye/

Hudon, C., Belleville, S., Souchay, C., Gély-Nargeot, M-. C., Chertkow, H., & Gauthier, S. (2006). Memory for gist and detail information in Alzheimer's disease and mild cognitive impairment. *Neuropsychology, 20*(5), 566–577. https://doi.org/10.1037/0894-4105.20.5.566

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2001). *Criminal interrogation and confessions* (4th ed.). Aspen Publishers.

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2013). *Criminal interrogation and confessions* (5th ed.). Jones & Bartlett Learning.

Innocence Project. (n.d.). *Eyewitness identification reform: Mistaken identifications are the leading factor in wrongful convictions*. Retrieved May 20, 2020, from https://www.innocenceproject.org/eyewitness-identification-reform/

Kandel, E. R., Dudai, Y., & Mayford, M. R. (2014). The molecular and systems biology of memory. *Cell, 157*(1), 163–186. https://doi.org/10.1016/j.cell.2014.03.001

Kassin, S. M. (2014). False confessions: Causes, consequences, and implications for reform. *Policy Insights from the Behavioral and Brain Sciences, 1*(1), 112–121. https://doi.org/10.1177/2372732214548678

Kassin, S. M., & Kiechel, K. L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science, 7*(3), 125–128. https://doi.org/10.1111/j.1467-9280.1996.tb00344.x

Kausler, D. H. (1991). *Experimental psychology, cognition, and human aging* (2nd ed.). Springer-Verlag Publishing.

Kircher, J. C. (2018). Ocular-motor deception test. In J. P. Rosenfeld (Ed.), *Detecting concealed information and deception: Recent developments* (pp. 187–212). Elsevier.

Kircher, J. C., & Raskin, D. C. (2016). Laboratory and field research on the ocular-motor deception test. *European Polygraph, 10*(4), 159–172. https://doi.org/10.1515/ep-2016-0021

Krokos, E., Plaisant, C., & Varshney, A. (2018). Virtual memory palaces: Immersion aids recall. *Virtual Reality, 23*(1), 1–15. https://doi.org/10.1007/s10055-018-0346-3

Leo, R. A. (2009). False confessions: Causes, consequences, and implications. Journal of the *American Academy of Psychiatry and the Law, 37*(3), 332–343.

Lipton, J. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology, 62*(1), 90–95. https://doi.org/10.1037/0021-9010.62.1.90

Loftus, E. F. (1976). Unconscious transference in eyewitness identification. *Law & Psychology Review, 2*, 93–98.

Løhre, E. (2011). *Context-dependent memory and mood* [Unpublished master's thesis]. University of Oslo.

Maki, R. H. (1990). Memory for script actions: Effects of relevance and detail expectancy. *Memory & Cognition, 18*(1), 5–14. https://doi.org/10.3758/BF03202640

Mapala, T., Warmelink, L., & Linkenauger, S. A. (2017). Jumping the gun: Faster response latencies to deceptive questions in a realistic scenario. *Psychonomic Bulletin & Review, 24*(4), 1350–1358. https://doi.org/10.3758/s13423-016-1218-z

McDaniel, M. A., Anderson, D. C., Einstein, G. O., & O'Halloran, C. M. (1989). Modulation of environmental reinstatement effects through encoding strategies. *The American Journal of Psychology, 102*(4), 523–548. https://doi.org/10.2307/1423306

Milne, B., Shaw, G., & Bull, R. (2007). Investigative interviewing: The role of research. In D. Carson, B. Milne, F. Pakes, K. Shalev, & A. Shawyer (Eds.), *Applying psychology to criminal justice* (pp. 65–80). John Wiley & Sons Ltd.

Morganti, F. (2004). Virtual interaction in cognitive neuropsychology. *Studies in Health Technology and Informatics, 99*, 55–70. https://doi.org/10.3233/978-1-60750-943-1-55

Murray, B. (2003). *The seven sins of memory: Convention award-winner Daniel Schacter explained the ways that memory tricks us*. American Psychological Association. https://www.apa.org/monitor/oct03/sins

Muzzio, I. A., Kentros, C., & Kandel, E. (2009). What is remembered? Role of attention on the encoding and retrieval of hippocampal representations. *The Journal of Physiology, 587*(Pt 12), 2837–2854. https://doi.org/10.1113/jphysiol.2009.172445

Norris, D. (2017). Short-term memory *and long-term memory are still different. Psychological Bulletin, 143*(9), 992–1009. https://doi.org/10.1037/bul0000108

Packard, M. G., Hirsh, R., & White, N. M. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *The Journal of Neuroscience, 9*(5), 1465–1472. https://doi.org/10.1523/JNEUROSCI.09-05-01465.1989

Patnaik, P. (2013). *Ocular-motor methods for detecting deception: Direct versus indirect interrogation* [Unpublished Master's thesis]. University of Utah.

Patnaik, P. (2015). *Oculomotor methods for detecting deception: Effects of practice feedback and blocking* [Unpublished Doctoral dissertation). University of Utah.

Patnaik, P., Woltz, D. J., Hacker, D. J., Cook, A. E., Ramm, M. L., Webb, A. K., & Kircher, J. C. (2016). Generalizability of an ocular-motor test for deception to a Mexican population. *International Journal of Applied Psychology, 6*(1), 1–9.

Perfect, T. J., & Harris, L. J. (2003). Adult age differences in unconscious transference: Source confusion or identity blending? *Memory & Cognition, 31*(4), 570–580. https://doi.org/10.3758/BF03196098

Read, J. D. (1994). Understanding bystander misidentifications: The role of familiarity and contextual knowledge. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), Adult eyewitness testimony: *Current trends and developments* (pp. 56–79). Cambridge University Press. https://doi.org/10.1017/CBO9780511759192.004

Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Christensen, A. (1990). The unconscious transference effect: Are innocent bystanders ever misidentified? *Applied Cognitive Psychology, 4*(1), 3–31. https://doi.org/10.1002/acp.2350040103

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*(1), 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Robin, J., & Moscovitch, M. (2013). The effects of spatial contextual familiarity on remembered scenes, episodic memories, and imagined future events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 459–475. https://doi.org/10.1037/a0034886

Ross, D. R., Ceci, S. J., Dunning, D., & Toglia, M. P. (1994). Unconscious transference and mistaken identity: When a witness misidentifies a familiar but innocent person. *Journal of Applied Psychology, 79*(6), 918–930. https://doi.org/10.1037/0021-9010.79.6.918

Schacter. D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*(3), 182–203. https://doi.org/10.1037/0003-066X.54.3.182

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Houghton Mifflin Company.

Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain. *Neuron, 20*(2), 185–195. https://doi.org/10.1016/s0896-6273(00)80448-1

Schuetzler, R. M. (2012). Countermeasures and eye tracking deception detection. *Information Systems and Quantitative Analysis Faculty Proceedings Presentations.* https://digitalcommons.unomaha.edu/cgi/viewcontent.cgi?article=1025&context=isqafacproc

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry, 20*(1), 11–21. https://doi.org/ 10.1136/jnnp.20.1.11

Sieberth, T., Dobay, A., Affolter, R., & Ebert, L. C. (2019). Applying virtual reality in forensics – A virtual scene walkthrough. *Forensic Science, Medicine and Pathology, 15*(1), 41–47. https://doi.org/10.1007/s12024-018-0058-8

Skowronski, J. J., Betz, A. L., Thompson, C. P., & Shannon, L. (1991). Social memory in everyday life: Recall of self-events and other-events. *Journal of Personality and Social Psychology, 60*(6), 831–843. https://doi.org/10.1037/0022-3514.60.6.831

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460–471. https://doi.org/10.1037/0278-7393.5.5.460

Smith, S. M. (1985). Background music and context-dependent memory. *The American Journal of Psychology, 98*(4), 591–603. https://doi.org/10.2307/1422512

Smith, S. M. (1986). Environmental context-dependent recognition memory using a short-term memory task for input. *Memory & Cognition, 14*(4), 347–354. https://doi.org/10.3758/BF03202513

Smith, S. M. (2007). Context and human memory. In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of Memory: Concepts* (pp. 111–114). Oxford University Press.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342–353. https://doi.org/10.3758/BF03197465

Smith, S. M., & Vela, E. (1992). Environmental context dependent eyewitness recognition. *Applied Cognitive Psychology, 6*(2), 125–139. https://doi.org/10.1002/acp.2350060204

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8*(2), 203–220. https://doi.org/10.3758/BF03196157

Snook, B., Luther, K., & Quinlan, H., & Milne, R. (2012). Let 'em talk!: A field study of police questioning practices of suspects and accused persons. *Criminal Justice and Behavior, 39*(10), 1328–1339. https://doi.org/10.1177/0093854812449216

Squire, L. R. (1987). *Memory and brain.* Oxford University Press.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99*(2), 195–231. https://doi.org/10.1037/0033-295x.99.2.195

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Squire. L. R. (2009). Memory and brain systems: 1969–2009. J*ournal of Neuroscience, 29*(41), 12711–12716. https://doi.org/10.1523/JNEUROSCI.3575-09.2009

Squire, L. R., & Dede, A. J. O. (2015). Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology, 7*(3), 1–14. https://doi.org/10.1101/cshperspect.a021667

Squire, L. R., & Kandel, E. R. (1999). *Memory: From mind to molecules*. W. H. Freeman and Company.

Squire, L. R., & Zola-Morgan, S. (1988). Memory: Brain systems and behavior. *Trends in Neurosciences, 11(*4), 170–175. https://doi.org/10.1016/0166-2236(88)90144-0

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science, 253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849

Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink. *Psychophysiology, 21*(1), 22–33. https://doi.org/10.1111/j.1469-8986.1984.tb02312.x

Thorley, C., Dewhurst, S. A., Abel, J. W., & Knott, L. M. (2016). Eyewitness memory: The impact of a negative mood during encoding and/or retrieval upon recall of a non-emotive event. *Memory, 24*(6), 838–852. https://doi.org/10.1080/09658211.2015.1058955

Tulving, E. (1979). Relation between encoding specificity and levels of processing. In L. S. Cermak, & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 394–418). Lawrence Erlbaum Associates.

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.

Vesisenaho, M., Juntunen, M., Häkkinen, P., Pöysä-Tarhonen, J., Fagerlund, J., Miakush, I., & Parviainen, T. (2019). Virtual reality in education: Focus on the role of emotions and physiological reactivity. *Journal of Virtual Worlds Research, 12*(1), 1–15. https://doi.org/10.4101/jvwr.v12i1.7329

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239–263. https://doi.org/10.1023/A:1006610329284

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10*(4), 141–142. https://doi.org/10.1016/j.tics.2006.02.003

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science, 20*(1), 28–32. https://doi.org/10.1177/0963721410391245

Webb, A. K, Honts, C. R., Kircher, J. C., Bernhardt, P. C., & Cook, A. E. (2009). Effectiveness of pupil diameter in a probable-lie comparison question test for deception. *Legal and Criminological Psychology, 14*(2), 279–292. https://doi.org/10.1348/135532508X398602

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45–75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Yu, L., & Gong, P. (2012). Google Earth as a virtual globe tool for Earth science applications at the global scale: Progress and perspectives. *International Journal of Remote Sensing, 33*(12), 3966–3986. https://doi.org/10.1080/01431161.2011.636081

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25–32. https://doi.org/10.1109/MC.2005.297

Appendix A

Pilot Study 1: Attention Task

Here is a list of objects which may have a sticker on them. Follow the order as listed, and check whether there is a sticker on each object. You may pick up the object but please place them back to their original location. You have 5 minutes to complete this task.

Is there a sticker on the object?

1. Water bottle: Yes No

2. Textbook: Yes No

3. Document tray: Yes No

4. File: Yes No

5. Potted plant: Yes No

6. Mug: Yes No

7. Calculator: Yes No

8. Waste bin: Yes No

Appendix B

Pilot Study 1: Free-choice Recognition Test

Not all of these objects were present in session 1. Please name the items you have previously seen in session 1.

- File
- Heart poster
- Mug
- Highlighter
- Scissors
- Waste bin
- Water bottle
- Textbook
- Stapler
- Calculator
- Hole puncher
- Floral poster
- Pencil holder
- Document tray
- Potted plant
- Photo frame

Appendix C

Pilot Study 1: Forced-choice Recognition Test

You will now see some pictures of the objects. You will be given 4 choices to select from. Please name the letter of the objects you have previously seen in session 1.

Appendix D

Pilot Study 1: Object Location Test

You will now see a picture of each object you have previously seen in session 1. Look around the room and you will see that number labels have replaced the objects you saw in session 1. Please name the correct location of each object.

1.		5.	
2.		6.	
3.		7.	
4.		8.	
9.		10.	

Appendix E

Study 2: Attention Task

Here is a list of objects which may have a sticker on them. Follow the order as listed, and check whether there is a sticker on each object. You may pick up the object but please place them back to their original location. You have 5 minutes to complete this task.

Is there a sticker on the object?

1. Headphones: Yes No
2. Bottled water: Yes No
3. File: Yes No
4. Mug: Yes No
5. Floral poster: Yes No
6. Textbook: Yes No
7. Duct tape: Yes No
8. Scissors: Yes No
9. Notebook: Yes No
10. Tea bag: Yes No
11. Heart poster: Yes No
12. Highlighter: Yes No
13. Calculator: Yes No
14. Pencil case: Yes No
15. Potted plant: Yes No
16. Hole puncher: Yes No
17. Stapler: Yes No
18. Thumb drive: Yes No
19. Mobile phone: Yes No
20. Ruler: Yes No
21. Pins: Yes No
22. Coffee sachet: Yes No

Appendix F

Study 2: Free-choice Recognition Test

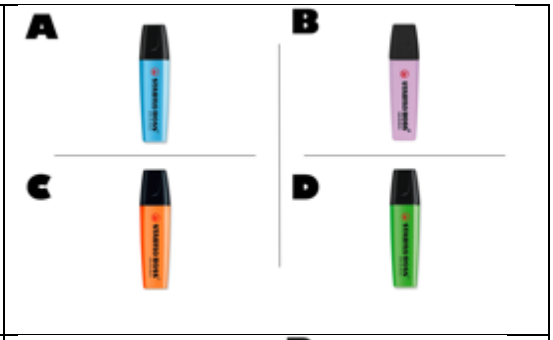
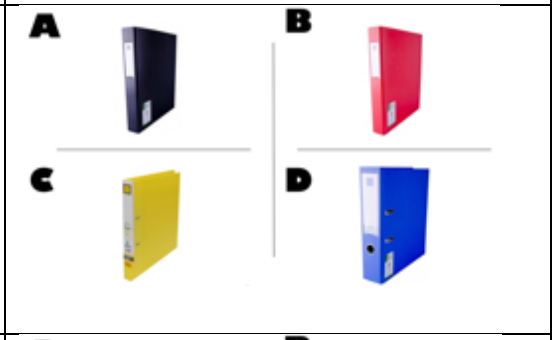
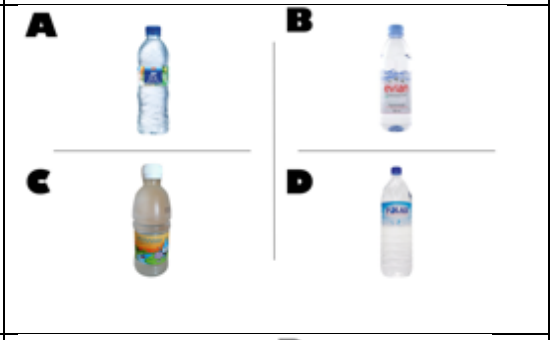
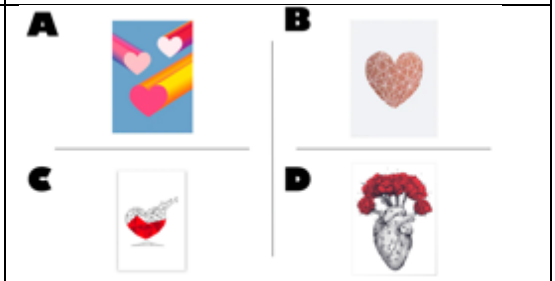
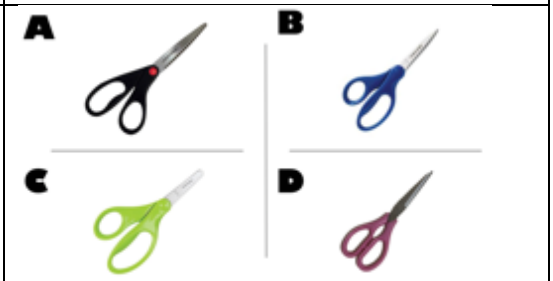
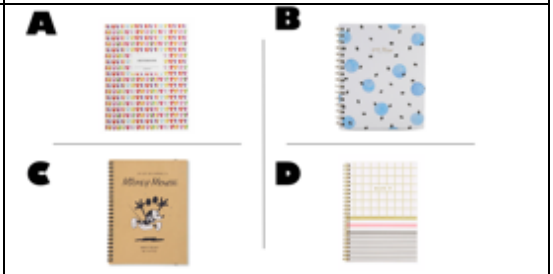
Not all of these objects were present in session 1. Please name the items you have previously seen in session 1.

- Pins
- File
- Heart poster
- Jacket
- Mug
- Umbrella
- Highlighter
- Bottled water
- Scissors
- Spectacles
- Slippers
- Notebook
- Pencil case
- Potted plant
- Pen knife
- Textbook
- Stapler
- Ruler
- Floral poster
- Headphones
- Photo frame
- Calculator
- Soft toy
- Tea bag
- Mobile phone
- Hole puncher
- Coffee sachet
- Keys
- Thumb drive
- Duct tape

Appendix G

Study 2: Forced-choice Recognition Test

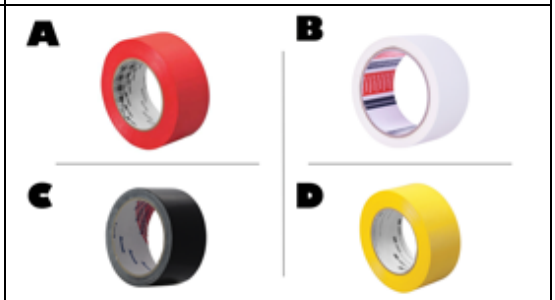
You will now see some pictures of the objects. You will be given 4 choices to select from. Please name the letter of the objects you have previously seen in session 1.

1.		5.	
2.		6.	
3.		7.	
4.		8.	

9.	A 	B 	
	C 	D 	
10.	A 	B 	
	C 	D 	
11.	A 	B 	
	C 	D 	
12.	A 	B 	
	C 	D 	
13.	A	B	
	C 	D 	

14.	A 	B 	
	C 	D 	
15.	A 	B 	
	C 	D 	
16.	A 	B 	
	C 	D 	
17.	A 	B 	
	C 	D 	
18.	A 	B 	
	C 	D 	

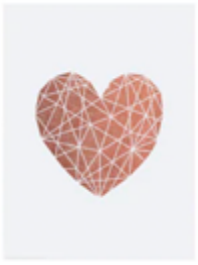
19.	
20.	

21.	
22.	

Appendix H

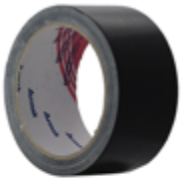
Study 2: Object Location Test

You will now see a picture of each object you have previously seen in session 1. Look around the room and you will see that number labels have replaced the objects you saw in session 1. Please name the correct location of each object.

1.		5.	
2.		6.	
3.		7.	
4.		8.	

9.		14.	
10.		15.	
11.		16.	
12.		17.	
13.		18.	

19.	
20.	

21.	
22.	

Accuracy Effects for ESS and Three-Position Scores of Federal ZCT Exams

Using the Grand Total Rule with Traditional/Federal and Multinomial Cutscores

Raymond Nelson

Abstract

This project is a comparison of accuracy effects for the ESS and three-position scores using traditional numerical cutscores and multinomial cutscores. Effects studied include test sensitivity, specificity, false-positive and false negative error rates, in addition to positive-predictive-value, negative-predictive value, the proportions of correct classification for guilty and innocent cases and the unweighted mean of correct and inconclusive cases. An archival samples of n=100 confirmed field cases using the Federal ZCT format were used, permitting intuitive comparison with previously published effects. A second sample of n=60 confirmed field cases using the Federal ZCT format was also included in the analysis. Responses were extracted from the recorded data and scores were assigned via an automated ESS algorithm that was designed to closely approximate the feature extraction process used by human experts when manual scoring polygraph data. ESS scores were then converted to three-position scores. A parametric bootstrap was used to calculate statistical confidence intervals at the .025 and .975 percentiles, and to estimate the variance of observed effects. A mixed-effects ANOVA procedure was used to evaluate the four treatments of the sample cases: ESS and three-position scores with traditional and multinomial cutscores. Accuracy for the four treatments when excluding inconclusive cases was similar for positive-predictive-value, negative-predictive-value, and the proportions of correct classifications excluding inconclusive results. Use of multinomial cutscores contributed to a statistically significant reduction of inconclusive results, and statistically significant increases in test sensitivity to deception and test specificity to truth-telling for both ESS and three-position scores. None of classifications of the individual cases were observed to change from deceptive to truthful or from truthful to deceptive for any of the four treatments with either of the two archival samples.

Introduction

The Empirical Scoring System (ESS; Nelson et al., 2011) is an evidence-based, standardized and statistically referenced method for the analysis of psychophysiological detection of deception (PDD) test data. The ESS can be thought of as a modification of the Federal three-position scoring method (National Center for Credibility Assessment, 2017), which can itself be thought of as a modification of the Federal seven-position scoring method. Previous studies on the ESS indicate that it provides accuracy effects similar to the Federal seven-position method with the practical reliability and intuition of the three-position scoring method. Results for the ESS were first described in a comparison of accuracy effects (Nelson, Krapohl & Handler, 2008) for polygraph examiner trainees with those from experienced examiners. The ESS was updated (Nelson, 2017a; 2017b) to make use of a Bayesian classifier and cutscores that were obtained from a multinomial reference distribution that was calculated under the analytic theory of PDD testing (Nelson, 2106). Multinomial cutscores and reference distributions were subsequently calculated for the three-position scoring method (Nelson, 2018a). ESS is widely used by polygraph examiners across the U.S. and worldwide, including professionals in private practice and in municipal, state and federal law enforcement/investigation agencies.

The ESS differs from the three-position scoring method in three main ways. First, ESS scores are assigned while doubling the value of all EDA scores. This is so that electrodermal scores are weighted in a manner that approximates the structural and statistical functions described in the scientific literature on PDD test data analysis and computer algorithm development Nelson (2019). A second important difference is that the ESS makes use of statistically referenced numerical cutscores in lieu of traditional cutscores that were derived heuristically for the seven-position scoring method. Another third difference is that different agencies have implemented the ESS with different decision rules, according to operational and mission objectives.

Decision rules define the structured procedures used to interpret and parse the categorical test result from numerical and statistical information. [Refer to Nelson (2018b) for a literature summary and description of polygraph decision rules.] Commonly used decision rules in PDD field practice include the grand-total-rule (GTR), two-stage-rule (TSR), and the Federal Zone Rule (FZR) and the subtotal-score-rule (SSR). Among these the GTR has been shown consistently to provide the highest level of classification accuracy for single issue exams, while the SSR has been regarded by many polygraph agencies and field examiners as the optimal decision rule for multiple issue screening exams.

This project is a comparison of accuracy effects for ESS using the GTR with traditional cutscores and multinomial cutscores. Two archival samples were used in this project, permitting intuitive comparison with previously published accuracy effects. Also studied were classification accuracy effects with three-position scores.

Methods and Materials

Data

Data for this project were a confirmed field sample of n=100 exams that were conducted using the Federal Zone Comparison Test (FZCT) format (Department of Defense, 2006). This sample was previously used by Krapohl and Cushman (2006) with Federal seven-position scores, and later by Nelson, Krapohl and Handler (2008) in an early study on the ESS. Sample cases were conducted by a variety of federal, state, and municipal law enforcement agency and were subsequently included in the confirmed case archive at the Department of Defense Polygraph Institute (now the National Center for Credibility Assessment). The FZCT is a three-question, event-specific test format, that is recognized as among the most useful test formats for the investigation of criminal incidents. All cases consisted of three iterations of a question sequence that included three relevant-questions (RQs) and three comparison-questions (CQs) in addition to other procedural questions that are not subject to numerical or statistical analysis. Field polygraph examiners refer to the repetitions or iterations of the question sequence as “charts,” with reference to old-time polygraph instruments that plotted physiological data through capillary ink pens onto rolled chart paper. Human expert, when scoring the sample cases manually, have described some of the sample cases as challenging. Although perhaps not ideal, use of this same sample data can provide practical and intuitive understanding of differences in accuracy effects for different test data analysis methods. [Refer to Nelson (2015) and Department of Defense (2006) for general information on the comparison question test and how the sample cases were conducted.]

All cases included the standardized array of PDD sensors, for which physiological responses and numerical scores would be extracted, including: upper and lower respiration sensors, an electrodermal activity sensor, and cardiovascular activity sensor. Acquired knowledge pertaining to the FZCT format, in addition to basic principles and procedures, have been generalized to other PDD formats including single-issue and multiple-issue use-cases with two, three and four RQs. This sample was used in the initial study and development of empirical reference distributions for the ESS, and was subsequently used in an accuracy demonstration of the multinomial update to the ESS-M (Nelson, 2017b).

A second archival sample was obtained, consisting of n=60 confirmed field exams using the FZCT format. These exams were also included in the DoDPI (now NCCA) confirmed case

archive. This sample was previously used as the holdout validation sample in the development of the OSS scoring algorithms (Krapohl & McManus, 1999, Krapohl, 2002, Nelson Krapohl & Handler, 2008), and was also used in a study of manual scoring with the Federal seven-position and ESS scoring methods. All exams in the second dataset consisted of three iterations of a question sequence that included three RQs and three CQs in addition to other procedural questions. Similar to the first archival sample, these examinations were conducted by a variety of municipal, state and federal law enforcement agencies.

Analysis

Sample data were analyzed using an automated version of the ESS. All tests data analysis methods – regardless of whether polygraph or other form of test – will consist of similar functions, feature extraction, numerical transformation and data reduction, use of some form of likelihood function or statistical classifier, and structured procedures for the interpretation and classification of result. Feature extraction refers to the identification of useful or diagnostic information in the recorded test data, and the extraction or separation of this information from other non-useful information or noise. Numerical transformation, when manually scoring polygraph test data, is the conversion of observed physiological responses to numerical values – using a system of [+, 0, -] integers. The simplest form of likelihood function is a numerical cutscore for which classification effects can be known, including true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) outcomes). Another form of likelihood function will map or obtain numerical cutscores to either an empirical or a theoretical reference distribution – both of which are available in publications for the ESS for which a multinomial reference distribution can be calculated under the analytic theory of PDD testing. Regardless of the form of likelihood function, parsing a categorical result from numerical and statistical test data requires the use of a structured decision rule. The automated ESS was designed to replicate objectively the procedures used by human experts when manually scoring PDD test data, including feature extraction, selection of RQ and CQ analysis spots, assignment and aggregation of integer scores, numerical cutscores, and decision rules.

Signal processing

Time-series data for all sample cases were exported to the NCCA ASCII format (Editorial Staff, 2019) and imported into the R Statistical Computing Language and Environment (R Core Team, 2019) to complete the signal processing, feature extraction, and data analysis. Signal processing of the digitized data was completed at a data rate of 30 cycles per second (cps) for all recorded sensors. Respiration data were subject to a smoothing filter, consistent with previous publications, consisting of a first-order Butterworth type low-pass filter (Butterworth, 1930) with a corner frequency of .886Hz (equivalent to a moving average filter with a .5 second window). Smoothing filters of this type have been shown to improve the correlation and diagnostic coefficients obtained from respiration data (Nelson & Handler, 2012).

All examinations were conducted using Axciton computerized polygraph systems that included a hardware-based high-pass filter (auto-centering EDA) option in addition to the manually-centered EDA option. Discussion with field practitioners revealed a common belief that field practices favored the use of manually centered EDA at the time the examinations were conducted. This may have been a result of the fact that engineering specifications of hardware-based high-pass filter of old-time analog polygraph instruments was largely undocumented as to the corner frequencies or time-constants of the filter design. Similarly, the corner frequency and time-constant for the Axciton computerized polygraph system has been described in previous publications as unknown (National Research Council, 2003). No information was captured or recorded regarding the selection of the EDA mode for the sample cases. A consequence of this is that it is possible that some of the sample cases were recorded using the hardware-based high-pass filter, and no attempt was made to determine the EDA mode through visual inspection. For this reason, no high-pass filter was used for the EDA data, and signal processing for EDA data was limited to the reduction of high-frequency noise through a first-order Butter-

worth type low-pass filter with a corner frequency of .886Hz.

Cardio data includes both low-frequency blood-pressure information and higher frequency pulse rate information. Because of the need to avoid altering or disrupting diastolic and systolic cardio peaks, cardio data was not subject to additional signal processing or smoothing.

The NCCA ASCII specification makes use of dimensionless units that are not associated with a standardized physical measurement. For this reason, scaling of the data has no effect on analytic results for individual cases or for this analysis.

Feature extraction

Feature extraction was accomplished using an automated procedure intended to replicate that used by human experts when manually scoring PDD data. Physiological reactions were evaluated using a 15 second evaluation window (EW) for all recording sensors. This EW is thought to be sufficient to observe most physiological responses to test stimuli and is regarded as somewhat robust for persons with common difficulties with sustained attention. For EDA and cardio data, a response-onset-window (ROW) was defined as from stimulus onset to five seconds after the point of verbal answer, or five seconds after stimulus offset if there was no recorded verbal answer.

Respiration feature extraction. For respiration data, information was excluded from the feature extraction for 1.5 seconds prior to and 1.5 seconds following the recorded point of verbal answer. This was to avoid the inclusion of commonly occurring answering distortions in the respiration feature extraction. Respiration data were measured using the respiration line excursion (RLE) – the sum of absolute change for each successive pair of respiration samples – using a sliding window of three seconds over the 15 second EW. For respiration rates in the normal range (10 to 22 cycles per minute) the sliding window would encompass ½ to 2 respiration cycles. The respiration measurement was the mean of all three second windows during the EW. For the 15 second EW at the 30 cps data rate, the feature extraction value was the means of the (15 – 3) * 30 = 360 three-second segments. Use of a sliding window in this manner means that feature extraction values are not dependent on the length of the EW and can be easily compared and optimized for different EW lengths – leading to potentially easier optimization of the EW. The response feature of interest is a reduction or suppression of respiration activity, that is expected to occur when a person attempts to conceal, or to avoid revealing or telegraphic, their deception. Although the automated feature extraction algorithm uses a dimensionless quantification of the RLE, human evaluators will observe respiration responses visually in plotted/displayed waveform patterns – as a subtle reduction of the respiration amplitude and as a subtle slowing of respiration rate.

EDA feature extraction. For EDA data, information was evaluated from stimulus onset to the end of the EW. Response peaks were identified as the change in EDA slope from positive (upward) to negative (downward) from 2.5 seconds after stimulus onset to the end of the EW. One additional response peak was also evaluated following the end of the EW if the EDA slope remained positive from 13.5 seconds to the first peak after the EW. This permits the extraction of information to the peak of response instead of the end of the somewhat arbitrary EW and also prevents the evaluation of a response peak after the EDA slope has turned negative late in the EW. Response onsets were identified as the onset of a positive slope segment (i.e., a change from negative or zero slope to positive slope) from a .5 second latency point (LP) to the end of the ROW. Use of the LP eliminates the need to evaluate the EDA slope prior to stimulus onset and ensures that responses that begin immediately with stimulus onset are not evaluated. When the EDA slope was already positive prior to stimulus onset and remained positive throughout the ROW a response onset was imputed statistically as a function of a statistically significant change in positive slope variance (with alpha = .001) for two adjacent one-second moving windows from the LP to the end of the ROW. This can be visualized by human experts as a substantial change in upward angle within a positive slope segment during the ROW. The extracted value was the maximum difference between a response peak and a preceding response onset. In simplistic terms the EDA response feature can be visualized as the max-

imum distance from a response onset during the ROW, after the LP, to a peak point during the EW.

Cardio feature extraction. Feature extraction for the cardio data was similar to that for EDA, but with two differences. Cardio data was extracted from a moving average of all recorded cardio data points. The moving average was calculated by passing the cardio data four times through a moving average filter of .5 seconds. The result of the moving average filter can be visualized as the mid-line between systolic and diastolic peaks. Inclusion of one addition response peak, after the 15 second EW, was retained if the EDA slope remained positive from 14.5 seconds to the first response peak after the EW. This change was needed improve the ability of the feature extraction procedure to tolerate the potential complexity of cardiovascular activity, which can sometimes be influenced by respiration activity in addition to cognitive, emotional and behavioral factors. Similar to the EDA feature, the cardio response feature can be visualized as the maximum distance from a response onset during the ROW to a peak point during the EW. One difference between the automated feature extraction and manual feature extraction is that human examiner will most often evaluate the cardio data at the diastolic baseline. This procedure is thought to improve the reliability of visual/manual feature extraction and is premised on a strong correlation between the information contained in the diastolic line and mid-line.

Numerical transformation and data reduction. All physiological responses were measured in dimensionless units – not intended to represent a physical quantity. This permits the scaling of data for visual display and plotting with no effect on the numerical transformations involving the comparison of RQ and CQ values. Data were assumed to be ordinal and intervalic. For EDA and cardio data greater extracted values were associated with greater changes in physiological activity. For respiration data the feature of interest to PDD testing is a reduction of respiration activity in respiration activity in response to RQs and CQs. This meant that smaller respiration values were associated with a greater change in physiological activity for the respiration scores.

Assumptions and constraints. Recorded physiological data and extracted values were not assumed to be linear, and no ratio assumptions were employed in the transformation of extracted response values to ESS integer values. However, some linear constraints were employed to prevent the extraction and scoring of extreme values. Extreme values were defined as less than 2% of the maximum scaling value for visual display and plotting. Values smaller than the 2% threshold were regarded as potential noise, and therefore less likely to be an authentic response to the test stimuli, and were not used in the transformation of extracted values to ESS scores. Leave-one-out z-scores were calculated for all extracted values for each recording sensor within each recorded test chart. For EDA and cardio data z-scores in excess of 10 (i.e., 10 standard deviations) were regarded as data artifacts, possibly resulting from physical movement, and were excluded from the analysis. None of the responses exceeeded this value ($z > 10$) for this sampling data, though other samples have included response artifacts in excess of 10 standard deviations. Extracted response values were assumed to be monotonic with changes in physiology. That is, greater changes in physiological activity were assumed to be associated with differences in the extracted numerical values.

ESS integer scores. ESS integer scores were assigned using a three position scale of signed values [+, 0, -] similar to the procedure for the three-position scoring method. One difference between ESS scores and three-position scores is that ESS scores are obtained using only the primary response feature whereas the three-position and seven-position methods permit the combined use of primary and secondary responses (National Center for Credibility Assessment, 2017).

ESS integer scores were assigned to each RQ after comparing after comparing the RQ with a paired CQ. RQ and CQ pairs were selected via automated algorithm using the procedure described by Nelson (2017c). FZCT cases in the sampling data consisted of three RQs – named R5, R7, and R10. For each recording sensor R5 is compared to either adjacent CQ (C4 preceding the RQ or C6 subsequent to the RQ) depending on which CQ has the greater change in physiological activity. Use of two CQs for R5

in this manner is thought to benefit innocent/truthful persons in that a change in physiological activity at the RQ will have to exceed that of two CQs before deceptive score can be assigned – also will also provide them with two opportunities to produce a change in physiological activity at a CQ that exceeds that of the RQ. R7 is compared only to the preceding CQ (C6), and R10 is also compared only with the preceding CQ. Field practices permit the rotation of some questions during the various iterations of the test question sequence; this is intended to distribute or balance effects related to the position of each question in the sequence and also to dissuade examinees from memorizing or habituating to the question sequence. Regardless of the rotation of questions, the first RQ in the sequence is compared to the first two CQs, while the second and third RQs are compared only to the preceding CQs. Each paired RQ and CQ is sometimes referred to by polygraph field examiners as an analysis spot.

Respiration constraints and numerical scores. For respiration data, ESS integer scores of + sign value, indicative of truth-telling, were assigned when a greater change in physiology was observed in response to the CQ, while scores of – sign value, indicative of deception, were assigned when a greater change in physiology was observed in response to the RQ. In contrast to the EDA and cardio data, greater changes physiology are observed in respiration data as smaller extracted values – indicative of a greater reduction or suppression of respiration activity. To prevent the analysis of extreme changes or extreme values that may result from voluntary or deliberate activity – such as that sometimes observed by persons attempting to alter or fake their test data and results – a maximum respiration constraint ratio of 1.5:1 was employed on the RQ and CQ analysis spots. This constraint is intended to prevent the assignment of a signed integer score when a takes a deep breath or holds their breath in response to an RQ or CQ. Additionally, a minimum respiration constraint ratio of 1.25:1 was used to prevent the assignment of numerical score to response differences that are not due to the test stimuli – and which may be considered noise resulting from either normal/uncontrolled variation in respiration activity, or due to observed instability that some persons exhibit in their respiration rate and amplitude.

ESS scores from the abdominal and thoracic respiration sensors are combined into a single ESS score using the procedure described by Nelson and Krapohl (2019). This procedure is common to other manual scoring methods and will be familiar to many field practitioners. Scores are combined to a value of zero (0) when the sign values are opposite for the thoracic and abdominal sensors, and are collapsed to a single singed score when they are not opposite.

EDA and cardio constraints and numerical scores. For EDA and cardio data, ESS integer scores of + sign value, indicative of truth-telling, were assigned when a greater change in physiological activity was observed at the CQ. Greater changes in physiology are observed in EDA and cardio data as larger extracted values. Scores of – sign value, indicative of deception, were assigned when a greater change in physiological activity was observed in response to the RQ. Scores of 0 (0 sign value) were assigned when there was no observable or appreciable difference between the responses to an RQ and CQ analysis spot. A minimum constraint was used to prevent the assignment of score to EDA and cardio for RQ and CQ analysis spots for which the observed difference in response magnitude was small. The constraint selected for this project was a ratio of 1.05:1, for RQ and CQ analysis spots. This constraint was the result of step-wise optimization of correlation and receiver operating characteristic (ROC) coefficients with other data. Differences smaller than 5% are more likely to be the result of physiological noise, and may also be the result of unknown influence on the EDA data when using an auto-centering EDA solution for which the design characteristics are unknown or undocumented.

Weighted EDA scores. ESS integer scores for EDA data are weighted more than for other recording sensors. This is accomplished by doubling all + and – integer values to +2 and -2. The effect of this is to approximate the structural and statistical coefficients that have been reported in numerous studies on PDD data analysis and computer algorithm development over a period of nearly five decades.

[Refer to Nelson (2019) for a literature survey on structural and statistical coefficients for respiration, EDA, cardio and vasomotor data.]

Data reduction. Because ESS scores are signed integer values, data reduction is a simple matter of summation. Subtotal scores are summed for each RQ, including all sensors and all iterations of the question sequence. Subtotal scores are then summed for a grand total score. Although field practices and other analyses may often make use of subtotal scores, this project involves the analysis of only the grand total score.

Multinomial likelihood function and numerical cutscores

Likelihood function. The simplest form of likelihood function is a numerical cutscore that correspond to a known empirical likelihood of a correct or incorrect classification. Formally, a likelihood function is a tool – including possibly a mathematical or statistical formula, computer function or published reference table – that can be used to calculate or obtain a statistical value for the observed test data. Cutscores for ESS scores can be obtained from multinomial reference tables and Bayesian analysis.

Both grand total cutscores and subtotal cutscores can be used to classify the test data as either indicative of deception or truth-telling – the contextual allegory of the more general terms positive and negative. Published studies have consistently shown that grand total scores provide the highest rates of classification accuracy. [Refer to APA (2011) for a summary of effect sizes for validated polygraph techniques.] This is not be surprising when considering that grand total scores make use of more information than the question subtotal scores and will therefore provide reduced variation and more opportunity for data to converge. This sometimes referred to as the weak law of large numbers (Dekking, 2005) – a related to the central limit theorem which states that the means of randomly selected samples will be normally distributed and will converge towards the unknown population mean. It is the main reason that it is advantageous to have many samples (for which meta-analysis can also be used) and the reason that larger samples are preferred over smaller samples.

The question of great importance is this question: what numerical cutscores are most effective or most efficient to classify test results as indicative of deception or truth-telling? Or, more precisely, what probabilistic inferences about deception and truth-telling can be made about the numerical cutscores and resulting classifications? Because scientific testing and scientific test data analysis is inherently probabilistic (given that the purpose of any scientific test is to quantify a phenomena of interest that cannot be subject to physical measurement), field examiners and program managers will be primarily interested in this more practical version of the same question: what numerical cutscores will provide an optimal experience of correct vs incorrect outcomes? In scientific terms this is the question of selecting numerical cutscores that will optimize the desired observation of TN, TP, FN, and FP results. In the polygraph context the answer to this question will be considered with regard to the additional outcome potential for inconclusive outcomes.

Analytic theory of polygraph testing. The analytic theory of polygraph testing – under which the multinomial distributions of ESS and three-position scores are calculated – holds that greater changes in physiology will be loaded at different types of test stimuli (i.e., relevant and comparison questions) as a function of deception or truth-telling in response to the relevant or target stimuli. [Refer to Nelson (2016) for a discussion of the analytic theory of the polygraph test.] In polygraph testing, some uncontrolled variation is expected at the level of each sensor and each presentation of each RQ (e.g., it is not expected that scores will be of uniform sign value). To the degree that the theory of PDD testing is valid (supported by evidence), and PDD sensors record valid data (data and scores that are loaded as a function of deception or truth-telling and not mere randomness) the convergence of subtotal and grand total scores can be used to make statistical inferences about reality – the degree to which a person is probably deceptive or probably truthful. In other words, it is the aggregation of subtotal and grand total scores that will be used to classify the test data as indicative of deception or indicative of truth-telling.

Aggregation of scores from multiple RQ, multi-

ple CQs, multiple recording sensors, and multiple iterations of the question relies on the law-of-large-numbers (LLN) – which for the aggregation of PDD scores will converge to become loaded to a value of either + or – sign value as a function of deception or truth-telling in response to the RQs. The LLN also provides insight as to why overall classification accuracy with grand total scores is expected to continue to outperform overall classification accuracy with subtotal scores.

Multinomial distribution is calculated under the analytic theory. The mathematical/statistical distribution of data values (i.e., all possible ESS scores and the probabilities associated with each) can be characterized empirically, by obtaining data from reality. A distribution of ESS scores can also be calculated using only information subject to mathematical and logical proof under a proposed theory. Mathematical characterization of a distribution of scores is often accomplished under the null-hypothesis to a theory. This is because it is often difficult (read: impossible) to mathematically characterize a proposed theory while the (null-hypothesis) can often be easily characterized as a distribution of random values. A well-known distribution is the Gaussian or normal (bell-curved) distribution. We use our mathematical knowledge of statistical distribution to make inferences about individual cases relative to the population of all possibilities that is represented by the statistical distribution. In the polygraph context, because there is a finite, though large, number of all possible combinations of ESS scores for all iterations of all questions and all sensors, the statistical distribution of ESS scores is not Gaussian, but is multinomial. The distribution of ESS scores is multinomial because there are three possible values for each score.

The multinomial distribution of ESS scores will exhibit a bell shape, somewhat similar to the normal distribution, though with discrete values for each possible test score. Under the null-hypotheses – that scores are not loaded in any systematic way and can therefore be characterized as random – most multinomial scores will occur near the middle of the distribution (near zero) with only one possible way to achieve the maximum or minimum score (uniform + or – scores for every iteration of every question and every sensor). There is a finite, though large, number of possible combinations of [+, 0, -] scores for each exam. There is also a finite number of ways to arrange the [+, 0, -] scores to achieve each possible score.

The multinomial distribution of scores is a list of all possible scores and the probability associated with each; it can be calculated using a combinatoric formula. It can also be calculated (sometimes more easily and quickly) via Monte-Carlo simulation. (Multinomial calculations during the Manhattan Project were an impetus for the development of Monte Carlo statistical methods.) The multinomial distributions for ESS and three-position scores (Nelson, 2017a; 2018a) are an exact calculation. Most importantly, our knowledge and information about the multinomial distribution of ESS scores can be used to make statistical inferences about reality (i.e., classifications under uncertainty). All that is necessary is to first calculate the likelihood statistic for an observed score, if loaded for deception or truth-telling, and then use the statistical value from the multinomial distribution as a likelihood function in Bayesian analysis of the likelihood of deception or truth-telling.

In addition to ESS scores, a multinomial distributions have also been published for three-position scores (Nelson, 2018a). This is possible because the three-position method relies on the bigger-is-better rule for which reactions that are recorded and measured, regardless of whether using standardized or dimensionless/arbitrary measurement units, are objectively either larger, smaller or equivalent. These differences are larger because there is mathematical proof that successive numbers, whether positive or negative, can in factual reality, represent in larger and smaller quantities – including when those quantities are not assigned a standardized measurement unit. For this reason, results in this analysis were also calculated for grand total scores of Federal three position scores. Unfortunately, no multinomial distribution exists for Federal seven-position scores – due to arbitrary decisions (i.e., without mathematical proof) as to the differences in physiological activity that correspond to the seven-position scale values. Automation of seven-position scores cannot be accomplished using only facts and informa-

tion subject to mathematical and logical proof, and for this reason questions about classification accuracy of seven-position grand total score were not addressed in this project.

Bayesian multinomial cutscores. Something that would be of great convenience would be to determine numerical cutscores that provide both a statistical classifier and also provide information about the practical meaning of the probabilistic strength of the classification. Multinomial cutscores for ESS cores (and three-position scores) together with Bayesian analysis do just that. Whereas early work on polygraph algorithms relied on statistical classifiers that were not intended to offer practical intuition or practical inference, multinomial cutscores, calculated using Bayesian methods, quantify both the practical or systematic likelihoods associated with deception or truth-telling in addition to the random error estimate associated with different outcomes that may result from the analysis of other data not available to the present analysis. Bayesian analysis is based on an assumption that the available sample/test data are all of the information available with which to make a conclusion (Stone, 2013; Winkler, 1972). In contrast, frequentist inference is based on an assumption that the available sample/test data are informative of the other data and information that could potentially be obtained from the universe and reality as it pertains to the individual and the behavioral target of a PDD investigation. [See Nelson (2017d) for a brief description of Bayesian analysis and null-hypothesis significance testing.] It is often the case the scientists and scientific methods may utilize a combination of frequentist and Bayesian assumptions. [Refer to Nelson (2018c) for a description of Bayesian analysis and the ESS-M.]

ESS Multinomial cutscores for grand total scores. For grand total scores with FZCT sample cases the multinomial grand total cutscores are - 3 or lower from deceptive classifications and +3 or greater for truthful classifications. There cutscores were selected from the multinomial distribution of all possible ESS scores (also the distribution of all possible ESS cutscores) at the point for which the random error estimate – indicated by the lower-limit of the Bayesian credible interval – provides a statistically significant likelihood (with alpha = .05) of continuing to observe the same analytic result, despite expected variation, if it were possible to repeat the examination or analysis numerous times. [Refer to Nelson (2018d) for a graphical illustration on the calculation of Bayesian ESS-M cutscores.]

Multinomial cutscores for three-position grand total scores. For three-position scores the multinomial cutscores can be calculated using the same Bayesian analytic methods as for the ESS. Multinomial cutscores for grand total scores of three-position scores are -2 or lower for deceptive classifications and +2 or greater for truthful classifications. [Also refer to Nelson (2020) for a tabular demonstration of ESS-M and three-position cutscores for a range of prior probabilities and different alpha levels for deceptive and truthful classifications.] Table 1 shows the multinomial cutscores for grand total scores with the three-position scoring method, along with the traditional cutscores for grand total scores.

Traditional cutscores. Traditional numerical cutscores were selected initially for older and more complex seven-position scoring methods; they too have been initially derived empirically and heuristically, and then subject to subsequent analysis for their classification efficiency. Traditional cutscores for grand total scores for FZCT exams are -6 or lower for deceptive classifications and +6 or greater for truthful classifications. An important consideration

Table 1. Traditional and multinomial cutscores for grand total scores with ESS and Federal 3-position scores.

	Traditional		Multinomial	
	Deception Indicated	No Deception Indicated	Deception Indicated	No Deception Indicated
ESS	-6	+6	-3	+3
Three-position	-6	+6	-2	+2

here is that field practice standards for Federal examiners who use the FZCT do not involve the use of grand total scores alone, and will instead involve a combination of grand total and subtotal scores. Although it is tempting to delve here and now into an empirical investigation of those procedures, and although little work has been published on the topic of decision rules since Senter and Dollins (2003), the purpose of this project was only to advance the available knowledge on effect sizes for numerical cutscores for grand total scores.

Another important consideration is that traditional grand total cutscores are also used with the three-position scoring method, leading to higher rates of inconclusive results for the three-position scoring method (APA, 2011) and the need to devote additional resources toward the resolution of these. Higher inconclusive rates also create a context for the emergence or reliance on covert solutions to reduce their occurrence. Most importantly, traditional cutscores were first suggested decades ago for the earlier and more complex seven-position scoring methods, and have remained unchanged despite scientific innovations in PDD data analysis and despite known and expected differences in the distribution of possible scores. Continued use of these traditional cutscores is a reflection of the fact that, although perhaps sub-optimal, outcome effects are reasonably known, and a more optimal solution, ideally supported by both theory and scientific evidence, has not yet been decided upon.

Interpretation and classification of analytic results.

Interpretation, in this usage, refers to the translation of numerical and statistical test results into categorical test results for which consistent and rational actionable decisions can be made. Interpretation and classification of test result is accomplished procedurally through the use of structured decision rules. Because this project involves the study of grand total cutscores, the decision rule of interest is the GTR.

Grand-total-rule. Execution of the GTR is a matter of summing the subtotal scores to obtain a grand total score. The grand total score is then compared to the numerical cutscores for grand total scores. Multinomial cutscores for the ESS and three-position methods are shown in Table 3. For ESS the multinomial cutscores are -3 or lower for deceptive classifications, and +3 or greater for truthful classifications. For three-position scores the multinomial cutscores are -2 or lower for deceptive classifications, and +2 or greater for truthful classifications. These cutscores are assuming a prior probability of 0.5. Traditional numerical cutscores for grand total scores are -6 or lower for deceptive classifications, and +6 or greater for truthful classifications. Traditional cutscores were derived for early seven-position scoring methods. Intuition suggests they may be inefficient for ESS and three-position scores – leading to higher rates of inconclusive results and over-reliance on subtotal scores. Although the use of subtotal scores, in addition to grand total scores, may improve classification with deceptive cases, this will introduce statistical multiplicity effects and may bias overall accuracy in unfortunate or unintended ways. For this reason, understanding and selection of optimal grand total cutscores may increase the accuracy effect sizes for the FZCT cases.

Bayesian analytic classification of deception or truth-telling. Multinomial grand total cutscores, for both ESS and three-position scores, provide a Bayesian posterior odds (systematic error) estimate of approximately 2:1 deception and truth-telling, permitting a 1-alpha x 100% = 95% likelihood of observing another analytic result of at least this value. In practical terms, test scores at this level are sufficient accept the notion that recorded physiological activity is loaded systematically, and to reject the notion that the scores are loaded in a random or meaningless un-interpretable/un-classifiable way. Although 2:1 odds may not be spectacular, it is important to recognize that classifications made at a score of +/- 2 or +/- 3 cannot, when considering the range of the distribution of possible scores, be reasonably expected to provide spectacular accuracy. Equally important, posterior odds of 2:1 may provide actionable knowledge for some circumstances. For example: consider the information that the odds of a particular bridge collapsing under weight are estimated at 2:1. Many reasonable persons might be quite hesitant to make use of that bridge. Of course, circumstances will also exist that may require a stronger basis of actionable probabilistic in-

formation than 2:1 posterior odds. For most estimated prior probabilities, these needs can be met via the multinomial reference data and the selection of numerical cutscores that will constrain systematic and random error rates to required levels.

Results

Classifications for deception and truth-telling were calculated for each of the two FZCT samples. Because polygraph field practitioners commonly discuss test accuracy effects in terms of the proportions of correct, incorrect and inconclusive conclusions, accuracy effects are presented in this way – instead of using effects sizes that compare classifications to chance levels.

Results for sample 1, n=100 confirmed FZCT field exams.

There were no cases that changed from positive to negative classification and no cases that changed from negative to positive classification as a result of the scoring method or type of cutscore for this sample (n=100) of confirmed field cases. Table 2 shows the test

Table 2. Grand total classifications for n=100 FZCT field sample with ESS and three-position scores

	ESS scores		Three-position scores	
	Traditional cutscores	Multinomial cutscores	Traditional cutscores	Multinomial cutscores
Error	[5] .05 (.02) {.01 to .10}	[5] .05 (.02) {.01 to .10}	[3] .03 (.02) {.01 to .07}	[5] .05 (.02) {.01 to .10}
Inconclusive	[35] .35 (.08) {.15 to .46}	[11] .11 (.05) {.01 to .18}	[53] .53 (.09) {.25 to .58}	[14] .14 (.05) {.01 to .21}
Correct	[60] .92 (.03) {.85 to .98}	[84] .94 (.02) {.89 to .99}	[44] .94 (.04) {.85 to .99}	[81] .94 (.03) {.88 to .99}
Sensitivity (TP)	[33] .66 (.07) {.53 to .79}	[44] .88 (.05) {.78 to .96}	[28] .56 (.07) {.43 to .70}	[43] .86 (.05) {.75 to .95}
Specificity (TN)	[27] .54 (.07) {.40 to .68}	[40] .80 (.06) {.69 to .91}	[16] .32 (.07) {.19 to .46}	[38] .76 (.06) {.64 to .87}
FN errors	[2] .04 (.03) {.01 to .10}	[2] .04 (.03) {.01 to .10}	[1] .02 (.02) {.01 to .07}	[2] .04 (.03) {.01 to .10}
FP errors	[3] .06 (.03) {.01 to .13}	[3] .06 (.03) {.01 to .13}	[2] .04 (.03) {.01 to .10}	[3] .06 (.03) {.01 to .13}
Inc guilty cases	[15] .30 (.06) {.18 to .43}	[4] .08 (.04) {.02 to .16}	[21] .42 (.07) {.29 to .55}	[5] .10 (.04) {.02 to .19}
Inc innocent cases	[20] .40 (.07) {.27 to .54}	[7] .14 (.05) {.05 to .24}	[32] .64 (.07) {.50 to .78}	[9] .18 (.05) {.08 to .29}
PPV	.92 (.05) {.82 to .99}	.94 (.04) {.86 to .99}	.93 (.05) {.83 to .99}	.93 (.04) {.86 to .99}
NPV	.93 (.05) {.83 to .99}	.95 (.03) {.88 to .99}	.94 (.06) {.80 to .99}	.95 (.03) {.88 to .99}
Correct guilty cases	.94 (.04) {.86 to .99}	.96 (.03) {.89 to .99}	.97 (.03) {.89 to .99}	.96 (.03) {.88 to .99}
Correct innocent cases	.90 (.06) {.78 to .99}	.93 (.04) {.85 to .99}	.89 (.08) {.72 to .99}	.93 (.04) {.84 to .99}
Unweighted inc.	.35 (.05) {.26 to .44}	.11 (.03) {.05 to .18}	.53 (.05) {.43 to .62}	.14 (.03) {.07 to .21}
Unweighted accuracy	.92 (.03) {.85 to .98}	.94 (.02) {.89 to .99}	.93 (.04) {.84 to .99}	.94 (.03) {.88 to .99}

* Cells show the [frequency] in addition to the bootstrap estimate of the mean, (standard deviation) and {95% confidence interval}.

accuracy metrics for classifications using the GTR with both traditional cutscores and multinomial cutscores for ESS and three-position scores. Included in Table 2 are the error and inconclusive rates, along with the proportion of correct classifications excluding inconclusive results. Also included in Table 2 are the sensitivity (TP) and specificity (TN) rates, along with FN and FP error rates. Other metrics in Table 2 are the positive predictive value (PPV) calculated as the proportion of true positive results and all positive results, and negative predictive value (NPV) which is the proportion of true negative and all negative classifications. Also shown are the proportion of correct decisions for guilty and innocent cases excluding inconclusive result, along with the unweighted accuracy and unweighted inconclusive rates.

Inspection of the rows in Table 2 indicates that the confidence intervals are substantially overlapping for the proportions of errors produced by the four treatments. However, some differences can be observed in sensitivity, specificity and inconclusive results. Both sensitivity to deception and specificity to truth-telling were greater for ESS scores and for multinomial cutscores. Inconclusive rates were lower for ESS scores and for multinomial cutscores. The frequency of TP and TN results was greater for the ESS and multinomial cutscores and lower for three-position and traditional cutscores. The frequencies of inconclusive results were higher for traditional cutscores and lower for multinomial cutscores.

Results for sample 2, n=60 confirmed FZCT field exams.

For the second sample of n=60 FZCT cases, there were no cases for which the classification changed from positive to negative or from negative to positive as a result of the scoring method or cutscore type.

Inspection of the rows in Table 3 indicates that results with the second FZCT sample paralleled those of the first sample. Confidence intervals are substantially overlapping for the accuracy metrics for correct classifications. Sensitivity to deception and specificity to truth-telling were greater for ESS scores and for multinomial cutscores. Inconclusive rates were lower for ESS scores and for multinomial cutscores.

Table 3 shows the same test accuracy metrics for ESS and Federal three position scores for the second archival sample.

Table 3. Grand total classifications for n=60 FZCT field sample with ESS and three-position scores*

	ESS scores		Three-position scores	
	Traditional cutscores	Multinomial cutscores	Traditional cutscores	Multinomial cutscores
Error	[2] .03 (.02) {.01 to .08}	[2] .03 (.02) {.01 to .08}	[1] .02 (.02) {.01 to .05}	[5] .08 (.04) {.02 to .15}
Inconclusive	[21] .35 (.10) {.10 to .50}	[7] .12 (.04) {.01 to .13}	[32] .53 (.11) {.22 to .63}	[6] .10 (.06) {.01 to .2}
Correct	[37] .95 (.04) {.87 to .99}	[51] .96 (.03) {.91 to .99}	[27] .96 (.04) {.89 to .99}	[49] .91 (.04) {.82 to .98}
Sensitivity (TP)	[21] .68 (.09) {.50 to .84}	[29] .94 (.05) {.83 to .99}	[17] .55 (.09) {.37 to .73}	[27] .90 (.05) {.79 to .99}
Specificity (TN)	[16] .53 (.09) {.34 to .72}	[22] .73 (.08) {.57 to .89}	[10] .33 (.09) {.17 to .50}	[22] .73 (.08) {.56 to .88}
FN errors	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[1] .03 (.03) {.01 to .11}
FP errors	[2] .07 (.05) {.01 to .17}	[2] .07 (.05) {.01 to .17}	[1] .03 (.03) {.01 to .11}	[4] .13 (.06) {.03 to .26}
Inc guilty cases	[9] .29 (.08) {.14 to .46}	[1] .03 (.03) {.01 to .11}	[13] .42 (.09) {.24 to .60}	[2] .07 (.05) {.01 to .17}
Inc innocent cases	[12] .40 (.09) {.23 to .59}	[6] .20 (.07) {.07 to .35}	[19] .64 (.09) {.45 to .81}	[4] .14 (.06) {.03 to .27}
PPV	.91 (.06) {.77 to .99}	.94 (.04) {.83 to .99}	.95 (.06) {.82 to .99}	.87 (.06) {.74 to .97}
NPV	.94 (.06) {.80 to .99}	.96 (.04) {.85 to .99}	.91 (.09) {.71 to .99}	.96 (.04) {.86 to .99}
Correct guilty cases	.96 (.05) {.85 to .99}	.97 (.03) {.88 to .99}	.95 (.05) {.81 to .99}	.96 (.03) {.88 to .99}
Correct innocent cases	.89 (.08) {.71 to .99}	.92 (.06) {.79 to .99}	.91 (.09) {.70 to .99}	.85 (.07) {.70 to .97}
Unweighted inc.	.35 (.06) {.23 to .47}	.12 (.04) {.05 to .20}	.53 (.06) {.40 to .65}	.10 (.04) {.03 to .18}
Unweighted accuracy	.92 (.05) {.82 to .99}	.94 (.03) {.87 to .99}	.93 (.05) {.81 to .99}	.91 (.04) {.82 to .98}

* Cells show the [frequency] in addition to the bootstrap estimate of the mean, (standard deviation) and {95% confidence interval}.

Analysis of the combined sample data

A two-way repeat measures ANOVA (scoring method x cutscore type) showed a significant interaction for inconclusive results [F (1,636) = 329.671, (p < .001)], indicating that observed differences in inconclusive rates for multinomial and traditional cutscores were different for ESS and three position scores. One way ANOVAs showed that the reduction of inconclusive results was statistically significant at the a=.05 level for ESS scores [F (1,318) = 4, (p = .046)], and also for the three position scores [F (1,318) = 12.308, (p < .001)].

A three-way repeat measures ANOVA for correct positive and negative classifications showed a significant three-way interaction, for criterion state (guilty vs innocence), scoring method (ESS, three-position) and cutscore type (traditional, multinomial) [F (1,632) = 54.282, (p < .001)]. Table 4 shows the ANOVA summary. All of the main effects and two-way interactions were also significant in the three-way ANOVA but were not interpretable due to the significant interaction effects.

Because differences in accuracy effects as function of cutscore type were the main interest for this project, a series of one-way contrasts was completed. For ESS scores the one-way effect was statistically significant for increased test sensitivity to deception [F (1,158) = 7.533, (p = .007)] and for increased test specificity to truth-telling [F (1,158) = 5.347, (p = .022)]. For three-position scores the one-way effect was also statistically significant for both increased test sensitivity to deception [F (1,158) = 11.148, (p = .001)] and test specificity to truth-telling [F (1,158) = 17.663, (p < .001)].

Risk ratios

To more adequately understand the differences in sensitivity, specificity and inconclusive rates shown by these data, risk-ratios were calculated after transforming the observed proportions to odds. Risk ratios are based on an assumption that observed proportions are an estimate of the likelihood or strength of the possibility of observing a similar outcome with any randomly selected member of the population, and are calculated as the ratio of the proportions observed for two different methods. In this project the comparison of interest is the risk-ratio for differences in outcomes for traditional and multinomial cutscores. Table 4 shows the risk-ratios for true-positive, true-negative and inconclusive results.

Risk ratios are informative as to the practical likelihood of differences in observed outcomes as they may be experienced for an individual or groups of cases. Risk ratios in Table 4 suggest that the use of multinomial cutscores with ESS scores may reduce the likelihood or occurrence of inconclusive outcomes to 34% of what would be expected from traditional cutscores. For three-position scores the risk of inconclusive outcomes may be reduced to approximately 20% of the risk of inconclusive results using traditional cutscores. However, actual rates in field practice the observed difference may not achieve these estimated differences because field practitioners may already engage in a variety of activities to resolve or reduce the occurrence of inconclusive results. Information shown in Table 4 indicates that guilty persons are 1.4 times more likely to be detected using multinomial cutscores, while innocent persons may be 1.6 times more likely to be classified as truthful.

Table 4. Risk-ratios for TP, TN, and inconclusive results for traditional and multinomial cutscore

	ESS scores	Three-position scores
Inconclusive	.34 (2.9)	.19 (5.3)
Sensitivity (TP)	1.38	1.38
Specificity (TN)	1.64	2.21

Conclusion

This project, concerned with the study of grand total cutscores, involved the use of the GTR with ESS and three-position scores using both traditional and multinomial cutscores. The GTR, although perhaps the simplest of all PDD decision rules, has been shown to provide the highest rates of overall classification accuracy among the variety of PDD decision rules. Accuracy effects were compared for ESS and three-position scores of event-specific polygraph exams using traditional numerical cutscores and multinomial cutscores for grand total scores. Although skill development at manual scoring with the ESS – as with PDD testing in general – is most effectively acquired as a function of both didactic or academic knowledge of standardized procedures and practical supervision and guidance under other experienced professionals, the basic concepts of the ESS are simple and highly structured, leading to the potential for an automated process that closely approximates the activities of human experts. During this project, to ensure that observed variance can be attributed to differences in numerical cutscores, and not due to expected variation within the inter-rater reliability limits of manual scoring methods, ESS and three-position scores and results were obtained via computer algorithm, including feature extraction, selection of RQ and CQ analysis spots, numerical transformation, data reduction application of the GTR with both multinomial and traditional cutscores.

Data for two different archival samples of confirmed FZCT field exams were used. These archival samples have been characterized by human scorers as challenging, though previously reported effect sizes for both human and computer algorithms is consistent with those shown herein. Results are shown separately in table form for each of the samples. Data for two samples were then combined for statistical analysis of potential differences in effect sizes for ESS and three-position scores. For the combined samples, significant differences were observed for test sensitivity to deception, test specificity to truth-telling, and the proportion of inconclusive results. Use of multinomial cutscores reduced the occurrence of inconclusive results and increased both sensitivity to deception and specificity to truth-telling. Use of archival data permits the direct comparison of observed effect sizes with previously reported effect using the same data with other scoring methods for which field examiners have intuitive knowledge and experience of their effectiveness.

Of particular interest in this project is that none of the deceptive or truthful classifications were reversed as a result of the selection of traditional or multinomial cutscores for grand total cutscores with either the ESS or three-position scores. Another interesting observation in this project is that accuracy effects for correct classifications, including PPV, NPV, and the proportion of correct classifications excluding inconclusive results within the guilty and innocent, along with the unweighted accuracy excluding inconclusive results were substantially similar for both traditional and multinomial cutscore with both ESS and three-position numerical scores. However, use of multinomial cutscores increased sensitivity to deception by factor of 1.4 for both ESS and three-position scores, while increasing specificity to truth-telling by a factor of 1.6 for ESS scores and by a factor of 2.2 for three-position scores. Multinomial cutscores reduced the occurrence of inconclusive results by a factor of 2.9 for ESS scores, and by a factor of 5.3 for three-position scores. Use of multinomial cutscores and grand total scores with the FZCT format, a three-question single issue format, achieves the level of accuracy requires by the American Polygraph Association Standards of Practice for evidentiary exams – those exams conducted with an expectation that the results and information may be used as information in a legal proceeding (American Polygraph Association, 2018)) – for both ESS and three-position scores.

Potential practical implications of these results include the possibility of increasing the effectiveness of field polygraphs in correctly classifying both deception and truth-telling. Another practical implication is that it is a reasonable consideration, in terms of classification accuracy, for polygraph programs to make use of traditional numerical cutscores with ESS scores if the prospect of changing both score type (ESS or three-position scores) and cutscores presents an uncomfortable number of degrees of freedom for policy mak-

ers. Indeed, there is practical and scientific wisdom in changing one variable at a time while observing and gaining experience with different methods. More broadly, these results support that it is a reasonable consideration for field examiners and/or policy makers to consider using only the grand total score for the FZCT format – a finding for which there is no basis of information or theoretical rationale to suggest that it can be generalized to all single issue polygraph formats.

Like all projects, this one is subject to some limitations. One difference between the procedures used during this project and those used by field practitioners when manual scoring is that this project is limited to the analysis of grand total scores. Human experts in field practice may rely on different decision rules depending on agency policy. Although not all agencies will choose to make use of only the grand total score, grand total scores have been shown consistently in published studies to provide the highest overall rates of classification accuracy. Use of grand total scores in this manner is thought to provide the greatest insight into the influence of grand total cutscores on overall test accuracy regardless of the decision rules used in field practice.

This project did not attempt to study effects with decision rules other than the GTR. Study of interaction effects involving both numerical cutscores and decision rules that may make use of both grand total and subtotal scores would require a multivariate analysis that is beyond the scope of this analysis – intended to be a simple an intuitive descriptive survey of accuracy effects with grand total scores. A more comprehensive project would have investigated both the type of cutscores and the decision rule. However, such a project would expand the complexity of the analysis considerably, along with a corresponding increase in the complexity of the analysis and information from the analysis. Limiting this project to 2 dimensions – scoring type and cutscore type – was thought to provide information of potentially practical use while also addressing the analytic questions in some degree of depth.

There is reason to expect that some interaction exists between the decision rule and the selection of numerical cutscores. One obvious implication of these analytic results is that traditional numerical cutscores for grand total scores, though not inaccurate, appear to be inefficient. A consequence of this is that polygraph field practitioners, in addition to polygraph trainers, quality control personnel and program managers, may have come to rely more heavily than is ideal on subtotal score to remediate the inefficiency. A consequence of over-reliance on subtotal scores to remediate inefficiency is that use of subtotal scores will introduce statistical multiplicity effects than complicate the test accuracy effects – most often in ways that can make the test appear biased against innocent persons. Selection of a more optimal grand total cutscore may increase test accuracy with both guilty and innocent persons while potentially relieving some of the burden of multiplicity effects. The interaction of decision rules and numerical cutscores should be the topic of further analysis and study.

Another limitation of this study involves the three-position scores. In this project three position scores were achieved by flattening of EDA scores of the ESS scores. It is unknown to what degree these three-position scores may differ from those achieved in field practice contexts where examiners might make use of secondary response features and other semi-subjective practices that are not included in the ESS and which cannot be subject to automation. In principle, three position scores can be extracted using the exact same automated procedure as for ESS scores. Three-position scores can also be achieved using the more complex system of primary and secondary features that was developed for seven-position scores – and which is less easily amenable to automation. Regardless of this limitation, it is the view of the author that the three-position values herein are sufficiently representative for these results to provide some potentially useful information.

A final limitation is that inconclusive rates observed in this study, like all scientific studies, may be greater than those observed in field practice. This to be expected. Polygraph field examiners, and polygraph programs, are regarded as acting reasonable if they engage in efforts to resolve inconclusive results at the level of each individual case. In research of this type such efforts would amount to manip-

ulating the research outcome. For this reason, no effort is made, in projects of this type, to resolve or reduce the occurrence of inconclusive results at the individual case level. Differences in inconclusive results are a reflection of the analysis method, and will necessarily be greater than inconclusive rates in field practice.

With consideration for the acknowledged limitations, accuracy effects observed in this analysis place the FZCT in the range required by the APA standards of practice for evidentiary examinations (APA, 2018). Evidentiary exams are those for which the test is conducted with the intention of introducing the test result as a basis of information in a legal proceeding. Accuracy rates observed herein equal or exceed those of other evidentiary PDD formats. Continued interest and continued research is recommended for ESS scores, the GTR and the use of multinomial cutscores.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. Polygraph, 40, 194-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph. org/section/research-standards-apa-publications.

American Polygraph Association (2018). APA Standards of Practice (Effective September 1, 2018). Retrieved (January 20, 2019) from https://www.polygraph.org/apa-bylaws-and-standards.

Butterworth, S. (1930). On the theory of amplifiers. Experimental Wireless and the Wireless Engineer, 7, 536-541.

Dekking, Michel (2005). A Modern Introduction to Probability and Statistics. Springer.

Department of Defense (2006). Federal psychophysiological detection of deception examiner handbook. Available from the author. Reprinted in Polygraph, 40 (1), 2-66.

Editorial Staff. (2019). Introduction to the NCCA ASCII Standard. Polygraph and Forensic Credibility Assessment, 48(2), 125-135.Polygraph & Forensic Credibility Assessment,

Krapohl, D. J. (2002). Short report: Update for the objective scoring system. Polygraph, 31, 298-302.

Krapohl, D. J. & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. Polygraph, 35(1), 55-63.

Krapohl, D. J. & McManus, B. (1999). An objective method for manually scoring polygraph data. Polygraph, 28, 209-222.

National Center for Credibility Assessment (2017). Test Data Analysis: Numerical Evaluation Scoring System Pamphlet. Available from the author. (Retrieved from http://www.antipolygraph.org on 4-13-2019).

National Research Council (2003). The Polygraph and Lie Detection. Washington, D.C.: National Academy of Sciences.

Nelson, R. (2015). Scientific basis for polygraph testing. Polygraph 41(1), 21-61.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. APA Magazine, 49(5), 69-82.

Nelson, R. (2017a). Multinomial reference distributions for the Empirical Scoring System. Polygraph and Forensic Credibility Assessment, 46 (2). 81-115.

Nelson, R. (2017b). Updated numerical distributions for the Empirical Scoring System: An accuracy demonstration with archival datasets with and without the vasomotor sensor. Polygraph and Forensic Credibility Assessment, 46 (2), 116-131.

Nelson, R (2017c). Heuristic principles to select comparison and relevant question pairs when scoring any CQT format. APA Magazine, 50(1), 73-83.

Nelson, R. (2017d). Five-minute science lesson: Bayesian and frequentist statistics – what's the deal? APA Magazine, 50(4), 89-95.

Nelson, R. (2018a). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. Polygraph and Forensic Credibility Assessment, 47(2), 158-175.

Nelson, R. (2018b). Practical polygraph: a survey and description of decision rules. APA Magazine, 51(2), 127-133.

Nelson, R. (2018c). Five minute science lesson: Bayes' theorem and Bayesian analysis. APA Magazine, 51(5), 65-78.

Nelson, R. (2018d). Practical polygraph: a tutorial (with graphics) on posterior results and credible intervals using the ESS-M Bayesian classifier. APA Magazine, 51(4), 66-87.

Nelson, R. (2019). Literature survey of structural weighting of polygraph signals: why double the EDA? Polygraph and Credibility Assessment, 48(2), 105-112.

Nelson, R. (2020). Multinomial cutscores for Bayesian analysis with ESS and three-position scores of comparison question polygraph tests. Polygraph and Forensic Credibility Assessment, 49(1), 61-72.

Nelson, R. & Krapohl, D. (2011). Criterion validity of the Empirical Scoring System with experienced examiners: Comparison with the seven-position evidentiary model using the Federal Zone Comparison Technique. Polygraph, 40, 79-85.

Nelson, R., Krapohl, D., & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. Polygraph, 37, 185-215.

Nelson, R. Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. Polygraph, 40, 67-78.

Nelson, R. & Krapohl, D. K. (2017) Practical polygraph: a recommendation for combining the upper and lower respiration data for a single respiration score. APA Magazine. 50(6), 31-41.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Senter, S. M. & Dollins, A. B. (2003). New Decision Rule Development: Exploration of a two-stage approach. Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph, 37(2), 149-164.

Stone, J. (2013). Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. Sebtel Press.

Winkler, R. L. (1972). An Introduction to Bayesian Inference and Decision. Holt Mc-Dougal.

Modification of the AFMGQT to Accommodate Single-Issue Screening: The British One-issue Screening Test

Donald J. Krapohl,

Donald Grubin,

Tim Benson,

and Bernard Morris

Abstract

There is a field requirement for a single-issue screening approach. The polygraph literature shows this need but has not explicitly described a screening technique which permits users to cover the same test question twice in the same test and to use the total score as a stand-alone primary decision rule. Here we discuss a method that is derived entirely from standard practices in the Federal You Phase and the Air Force Modified General Question Technique. Those practices, when combined, produce a useful single-issue screening methodology. Validation research summarized in the 2011 *APA Meta-Analytic Survey of Criterion Accuracy of Validated Polygraph Techniques generalizes well to this new method.*

Introduction

The polygraph has been in the public consciousness for nearly 100 years, beginning with newspaper accounts of crimes being solved by the then-new scientific tool for law enforcement. Since then the common depiction of the polygraph has been in its crime-solving role or in other settings to investigate unresolved suspicions. What may be surprising to many people is that today the main application of polygraph testing, at least in terms of sheer numbers, is for screening rather than in criminal investigations. Screening polygraphs are those conducted for the purpose of identifying possible problems, in the absence of a known allegation or known incident. Polygraph screening is now well established in settings such as police candidate selection and government security clearance vetting in scores of

[1]APA Past President and regular contributor to this publication. He is the Director of the BMUK Polygraph Training Centre.

[2]Emeritus Professor of Forensic Psychiatry at Newcastle University and CEO of Behavioural Measures UK.

[3]Police polygraph examiner, Hertfordshire, UK.

[4]HM Prison & Probation Service polygraph examiner, Oxfordshire, UK.

The authors are grateful for the helpful suggestions of Donnie Dutton in this article. The authors also appreciate the assistance of the APA Editor Mark Handler and the anonymous reviewer.

The views expressed in this article are those of the authors and do not necessarily represent those of the APA, the UK government, or the organizations with which the authors are affiliated. None of the authors have any financial interests in the method presented in this paper. The authors authorize the reprinting of this article for training purposes by APA accredited polygraph education programs. Questions and comments should be directed to the first author at APAkrapohl@gmail.com.

countries, as well as for monitoring sex offenders in the community, to name a few. Lesser known polygraph screening applications include the assessment of clandestine reporting sources, pursuit of concealed financial assets, and verification of rule compliance in sporting competitions.

While to the lay public a polygraph test is a polygraph test, there are distinctions between criminal and screening polygraph tests in that the protocols differ from one another in important ways, differences that result from the reasons the examinations are conducted. In criminal testing the agency conducting the examination knows the what (e.g., bank robbery, shooting, destruction of property, theft, etc.) that forms the basis of the test and needs to know who did it. In screening the organization has the who in the polygraph test chair but needs to know what the examinee may have done. In the former case the test questions focus on the specific event under investigation. A correct decision of deception will tell investigators the who (including who it isn't). In the latter case the examination covers a range of topics, the truth about which the agency considers important to a decision process, but where nothing may have occurred; in many instances the examinee might be the only person who knows the truth about these topics as it relates to him or her. Furthermore, in polygraph screening a deceptive decision will only reveal what the problematic topic is, but not what the examinee has actually done. Getting to those details requires elicitation to encourage self-report from the examinee.

In screening, examinee self-report of behaviors of interest to the agency becomes one of its valued features. This is as true for the use of the polygraph by police agencies (Meesig & Horvath, 1993) as it is in the management of convicted sex offenders (Gannon, Wood, Pina, Tyler, Barnoux & Vasquez, 2014; Grubin, 2008; Wood, Alleyne, Ciardha & Gannon, 2020). Self-report is similarly essential for security clearance determinations. According to the US Department of Defense Personnel Security Research Center (PERSEREC, 2000), "…over 95% of the information the NSA (National Security Agency) develops on individuals who do not meet federal security clearance guidelines is derived via voluntary admissions from the polygraph process" (parenthesis added).

A second product of polygraph screening are test results regarding the veracity of the examinee, which is complementary to disclosures. The validity of those test results depends upon the use of valid testing and analysis methodologies. In 2011 the American Polygraph Association (APA) undertook a sweeping assessment of field polygraph techniques to produce a report identifying those that could be supported by replicated and published empirical research. The report listed only two polygraph techniques generally used for screening that were validated, the Directed Lie Screening Test (DLST) and the Air Force Modified General Question Technique (AFMGQT). Both methods produced correct decisions at a rate greater than 85% when the Empirical Scoring System (ESS) was used as the method of data analysis (Blalock, Cushman and Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Nelson, Krapohl & Handler, 2008).

The DLST and the AFMGQT are both used for mixed-issue screening. The DLST is designed to test two separate issues in each series, while the AFMGQT is more flexible and can accommodate two, three or four issues. Both are in wide practice. Among police and probation polygraph examiners in the UK who conduct pre- and post-conviction sex offender testing the AFMGQT is the technique of choice for almost all screening examinations.

British government and police examiners are required to abide by APA Standards of Practice including the requirement to use validated techniques. They encountered a problem, however, that is not addressed in the APA literature: What method can be used when the

[5]As the 2011 APA report states, the report was being finalized while the APA Board was considering a change to its Standards of Practice that would permit the use of any screening technique that performed significantly greater than chance rather than the 80% minimum that had been applied to the techniques in the 2011 report. The change affected only the Relevant-Irrelevant Screening Test. It was subsequently approved. In recent years blind scoring of confirmed field test charts conducted with this technique have not shown its accuracy to be competitive with the other two screening techniques. It is not considered further in this paper.

requestor for the exam needs the examinee tested on only one screening issue? Neither the DLST nor the AFMGQT were intended to test just one screening issue by itself, nor does any of the validity research show that either method has been validated for just such an instance. As such there did not appear to be any option for polygraph examiners who have screening cases with just a single issue to resolve. This was the case, however, in about 10% of all sex offender testing conducted by UK police and probation examiners.

As an immediate remedy the examiners were permitted to use the Federal You Phase format, which is the only validated two-question technique that tests the same test question twice without including questions about other issues. The advantages of this method are that it allows two relevant questions to cover the exact same behaviors, for the scores for both questions to be added together to form a decision, and there is already published validity research for it. In time it became apparent, however, that the use of the You Phase in this manner did not come problem-free. First, the rules of the You Phase do not permit relevant question rotation. Rotating the relevant questions is intended to balance out position effects (e.g., habituation) that may influence the examinee's responses. Second, the You Phase also has symptomatic questions which have not been shown to function as intended (Cushman & Krapohl, 2010; Nelson, Handler, Oelrich & Cushman, 2014): anecdotal reports from polygraph examiners testing outside the US suggest symptomatic questions have often proven problematic with other cultures. Finally, because the You Phase is regarded as a technique exclusively for specific-issue testing, at least in all journal publications we were able to locate, it placed the UK government and police polygraph examiners in a position of being the only ones to use the method as a single issue screening test. It risked creating confusion between the UK examiners and all other examiners who use the You Phase as a specific-issue or diagnostic test.

The alternative would be to use the AFMGQT. However, the AFMGQT as commonly taught does not permit the same question to be tested more than once in a series. Standard rules also deny examiners the option of using grand totals as a decision rule as the Federal You Phase does. As such, the AFMGQT did not present itself as the needed solution.

There is a third approach. By simply changing two conventional rules of the AFMGQT it would be possible to address the field requirement. Those rules are to allow examiners to test the same question twice in the same test, and to use total scores as part of the decision rules. In a similar vein, Nelson, Handler, Oelrich and Cushman (2014) described the use of the AFMGQT as an event-specific or single-issue exam which would allow all scores to be totaled. However, they did not propose its use in screening. Because the method described here departs from the conventional AFMGQT, it was important to give it another name to distinguish it from that method. More about that later in this article.

While asking the same question more than once in a series is scarcely reported in the screening literature it is practiced without exception in specific-issue testing. The same rationale supporting the use of two versions of the same question in specific-issue testing can be extended to screening exams.

Naming

Because a description of a single-issue screening technique is not found in the literature, especially any that use precisely our combination of testing and analytical approaches, we felt compelled to assign a name to it. A unique name would help bring specificity to this particular approach and avoid confusion among others who may interpret the technique at first impression as a conventional AFMGQT. We resisted the temptation to name it after ourselves or to give it one built upon the name of an existing method as is often tradition[6]. Be-

[6] The naming history of the AFMGQT is an example of the unintended consequences that can accompany this tradition. The General Question Technique (GQT) was taught at the US government school by the 1960s. In the late 1960s the school director, Mr. Ron Decker, made changes to it so that it approximated the Reid Technique except that it included manual scoring. As the story goes, Mr. Decker approached John Reid to request permission to extend the title "Reid Technique" to

cause the technique is for screening on one issue and was devised in Great Britain, we took the easiest route and dubbed it simply the British One-issue Screening Test, or BOST

Description of the BOST

The BOST testing format is identical to the two-relevant question screening AFMGQT, variation 1 (Krapohl & Shaw, 2015; Department of Defense, 2006). The sequence of the test question presentation is shown in Figure 1.

Figure 1. Test question sequence for the BOST.

Position						
1	2	3	4	5	6	7
Neutral	Sacrifice Relevant	Comparison	Relevant	Comparison	Relevant	Comparison
N1	SR2	C3	R4	C5	R6	C7

Both relevant and comparison questions are rotated in this technique, but question types remain in the positions shown in Figure 1. Neutral questions can be inserted into the sequence as needed to re-stabilize the recordings after an examinee movement or prolonged physiological response. The principles for test question construction and presentation follow those of Krapohl and Dutton (2020). These principles include the use of words designed to prompt memories on the relevant questions for deceptive examinees, avoidance of evocative terms, construction of comparison questions that exploit cognitive load, doubt or examinee deception, and management of potential confounds in test question construction and presentation.

Relevant Questions

In the BOST the two relevant questions must cover the same behavior (that is, frame of reference) and time period as one another. For example:

R1. In the last six months have you entered your exclusion zones?

R2. Since last (month that precedes the current date by six months, e.g., September), have you gone into your exclusion zones?

Other question prefixes may include: since your last test, since the last time we met, since coming out of prison, since moving into the hostel, etc. These prefixes will be different between the two relevant questions, but both must cover precisely the same period.

In the UK the BOST is not used to combine several issues into a single test question, for example, in relation to denying the breach of two or more conditions of parole, or testing on a written statement using a single test question. It would likewise be incorrect to have the test question query whether the examinee is in breach of his parole conditions in general. Rather, the approved approach is to use behaviorally descriptive relevant questions.

Comparison Questions

Both Directed-Lie and Probable-Lie Comparison Questions are permitted with the BOST. The decision on which type is chosen relies on

this modified version. John Reid objected, saying in effect that he didn't care what the government school called it, just don't call it the Reid Technique. Ron Decker then named it the Modified General Question Technique, adding one more letter to the GQT acronym to create the MGQT. In the mid-1970s the US Air Force created a method that was purportedly a variation of the MGQT and placed their initials at the beginning of the acronym, now making it the AFMGQT. Being reluctant to break tradition, but mindful that the acronym was already a bit cumbersome, we considered then rejected calling our method the Single-Issue Screening AFMGQT with the acronym SISAFMGQT. We trust history will understand.

the judgment of the testing examiner. Whichever type the examiner employs, all comparison questions in the test must be of that same type.

Sacrifice Relevant Question

The BOST uses a sacrifice relevant question in the second position as is standard practice with the AFMGQT. It is phrased to be answered "yes." The standard verbiage is: Regarding _____ do you intend to answer truthfully? Other phrasing is permissible so long as the question addresses whether the examinee intends to answer truthfully specifically to the behavior in the relevant questions. These questions are included in the sequence based on their face-validity rather than any compelling evidence they serve a specific function.

Neutral Questions

All polygraph techniques utilize neutral questions, variously called irrelevant or norm questions. Neutral questions are intended to serve several functions such as permitting the physiological data to stabilize after an artifact or strong reaction, to record samples of baseline physiology, and in most techniques to be the first question in the test sequence. Neutral questions are selected and presented in the traditional way in the BOST. Examiners are encouraged to review two or more additional neutral questions with the examinee in case they are needed.

Scoring

UK government and police examiners have adopted ESS as the official scoring system. Because the BOST test questions cover identical issues, scores can be totaled across both test questions. The BOST also uses two-stage rules (Senter, 2003; Senter & Dollins, 2008) such that if the test results are Inconclusive (INC) with the total scores the examiner turns next to the sums of scores for each of the two relevant questions. Decision rules are guided by policy, which means endusers may establish their own cutoff scores based on risk tolerance. At this writing the UK government and police examiners are using +2 and -4 total scores. When there is a second stage, the cutoff score is -6 for the total of either relevant question.

The BOST also allows for the 3 – 5 chart decision rules. That is, if the results are INC after three charts, examiners should record another chart with the same questions again. If the four charts likewise result in an INC, a fifth and final chart can be recorded. No more than five charts are permitted. The cutoff scores remain the same regardless of the number of charts recorded.

Validation

The BOST was devised using components from both the Federal You Phase and the AFMGQT – its development did not entail the creation of any unique or novel processes. From the You Phase comes the repetition of the same relevant question within a test, not testing on secondary issues, and using the combined scores as part of its decision rules. From the AFMGQT the BOST has adopted all its testing protocol, including question rotation. The BOST has no practices that fall outside of either of these techniques. Because the practices of the BOST are constrained between the AFMGQT and the You Phase we argue that a separate series of validation studies are unnecessary because the BOST is wholly derived from polygraph techniques that have already undergone that scientific scrutiny. The APA meta-analysis (2011) reported the You Phase and AFMGQT have comparable accuracies between 87% and 90% when scored with ESS.

Conclusion

To meet a field requirement for a single-issue screening technique we combined components of two validated polygraph techniques and developed a third method called the BOST. It is not a significant departure from existing practices, though this is the first published report of those practices in this configuration. There is every reason to believe the BOST, or a similar approach, can already be found in the field despite not being documented in the polygraph literature.

The BOST takes advantage of the strengths of the two techniques from which it evolved, the You Phase and the AFMGQT. The use of two presentations of a screening question in a test, along with combining their scores, can be expected to provide a stable estimate of the ex-

aminee's veracity as it does in the You Phase. Indeed, with six presentations of the relevant question in three charts it is reasonable to believe that the BOST will produce greater accuracy than mixed-issue screening techniques that normally have only three presentations in the same number of charts.

Before concluding, we wish to reiterate the BOST is not a new format but a combination of existing field practices. It only represents a unique configuration of procedures not found in the existing polygraph literature for single-issue screening. Its naming is intended to form an operational distinction between it and the traditional AFMGQT and Federal You Phase approaches. It is the first report of a methodology to fill this field requirement. We do not suggest others are not using a similar method, but only no one has published that protocol or its rationale. The widely used rule set for the conduct of the AFMGQT means that users needing a single-issue screening technique can quickly adapt to the BOST. We suggest examiners may find it of benefit to add the BOST to their inventory of techniques for circumstances in which it is appropriate.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. Polygraph, 40(4), 196 – 305.

Blalock, B., Cushman, B., and Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. Polygraph, 38(4), 281 – 288.

Cushman, B., and Krapohl, D.J. (2010). The Evidence for Technical Questions in Polygraph Techniques. Presentation at the American Polygraph Association Annual Seminar, Myrtle Beach, SC.

Department of Defense (2006). Federal psychophysiological detection of deception examiner handbook. Reprinted in Polygraph, 40(1), 2-66.

Gannon, T. A., Wood, J. L., Pina, A., Tyler, N., Barnoux, M. F. L., & Vasquez, E. A. (2014). An evaluation of mandatory polygraph testing for sexual offenders in the United Kingdom. Sexual Abuse, 26, 178–203.

Grubin, D. (2008). The case for polygraph testing of sex offenders. Legal and Criminological Psychology, 13, 177–189.

Krapohl, D.J., and Dutton, D.W. (2020). A proposed framework for polygraph test questions. Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice, 49(1), 24 – 34.

Krapohl, D.J., and Shaw, P. (2015). Fundamentals of Polygraph Practice. Academic Press: San Diego, CA.

Meesig, R.T., and Horvath, F. (1993). Changes in usage, practices and policies in pre-employment polygraph testing in law enforcement agencies in the United States: 1964 – 1991. Polygraph, 22(1), 1 – 16.

Nelson, R. Handler, M. Oelrich, M. & Cushman, M. (2014). APA research committee report: proposed usage for an event-specific AFMGQT test format. Polygraph, 43(4), 156-167.

Nelson, R., Krapohl, D. J. & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. Polygraph, 37, 185 – 215.

Perserec (2000, Nov). Security Clearances and the Protection of National Security Information: Law and Procedures. Technical Report 00-4. Monterey, CA.

Raskin, D.C., Kircher, J.C., Honts, C.R., and Horowitz, S.W. (1988, May). A Study of the Validity of Polygraph Examinations in Criminal Investigations. Final Report to the National Institute of Justice. Grant No. 85-IJ-CX-0040. University of Utah, Salt Lake City, UT.

Senter, S.M. (2003). Modified General Question Test rule exploration. Polygraph, 32(4), 251 – 263.

Senter, S.M., and Dollins, A.B. (2008). Exploration of a two-stage approach. Polygraph, 37(2), 149 – 164.

Wood, J.L., Alleyne, E., Ciardha, C.Ó., and Gannon, T.A. (2020). An Evaluation of Polygraph Testing by Police to Manage Individuals Convicted or Suspected of Sexual Offending. University of Kent, Centre of Research and Education in Forensic Psychology (CORE-FP).

Timely Non-deceptive Sexual History Polygraph Examinations are Correlated with Completion of Treatment but Not Correlated with Sexual Recidivism

James E. Konopasek

Johneen Manno

Abstract

This research examines the speed at which individuals convicted of sexual offenses pass a sexual history polygraph examination (SHPE) and whether timeliness of achieving a truthful polygraph result is correlated with sexual recidivism. Findings indicate that SHPE results are positively associated with treatment completion in this sample, but not associated with sexual recidivism. Utilizing a convenience sample of 280 convicted sexual offenders, who were being treated in programs that were directed/administrated by the authors, this study refines and augments prior research (Konopasek, 2011; Konopasek, 2015; Konopasek & Nelson, 2015). Though findings may be informative to treatment programs utilizing sexual history polygraph testing as a treatment milestone, and may be relevant to supervision officers, passing a SHPE was not statistically associated with sexual recidivism in this sample, regardless of whether or not the polygraph examination was passed during the early months of treatment.

Keywords *sexual offender recidivism, community supervision, treatment outcome, Static-99R, polygraph, sexual history disclosure, timely sexual history polygraph examination*

Since 2010, the lead investigator of this study has hypothesized that sexual offenders who pass sexual history polygraph examinations (SHPEs) in a timely manner have better treatment outcomes and lower probability of sexual recidivism than sexual offenders who never pass a sexual history examination, or do not pass a SHPE in a manner deemed timely by treatment and supervision personnel. Research findings have been mixed relative to whether passing a SHPE (in a timely manner or not) has any correlation with recidivism (Cook, 2011; Cook, Barkley & Anderson, 2014; Konopasek, 2015; Konopasek & Nelson, 2015). Other researchers have focused on the question of whether post-conviction sex offender testing (PCSOT), utilized for sex offender compliance monitoring, has any correlation with recidivism. These studies have also produced mixed results (Abrams & Ogard, 1986; McGrath, Cumming, Hoke & Bonn-Miller, 2007). Rosky (2012) conducted a comprehensive review of the studies on PCSOT as related to deterrence, questioning the utility of PCSOT, and concluding that such endeavors are futile. The current study is intended to further explore the notion that timeliness of passing a SHPE may provide some predictive utility relative to sexual recidivism.

In his early research on this topic (Konopasek, 2011), utilizing bivariate analysis and logistic regression, found no statistically significant associations between passing a SHPE within 12 months of treatment onset, or ever passing SHPE, with the variable of sexual recidi-

vism. In the sample of 192 sexual offenders (which included males, females and remanded juveniles), the only variables found to be correlated with sexual recidivism were: a) age that truthful SHPE polygraph was obtained – over/under age 35 ($r\Phi$ = .151, p = .013); b) presence of sexual deviance -- as documented in penile plethysmograph and sexual interest viewing time examination reports ($r\Phi$ = .146, p = .044); and c) Static-99R risk score (r = .355, p = .030).

In follow-up research that excluded females from the analysis because the Static-99R was normed for assessing the risk of adult male sexual offenders (Konopasek, 2015) found that the variables of expeditious sexual history disclosure supported by a non-deceptive SHPE (i.e., passing a SHPE within 12 months of treatment onset) and ever passing a SHPE, were not correlated with sexual recidivism. Findings that the variable of passing a SHPE was correlated with the variable of treatment completion were not meaningful because passing a SHPE was considered a treatment milestone for the sexual offenders examined in the study – and therefore the variables were not independent.

In similar research (Konopasek & Nelson, 2015) examined the same variables evaluated in Konopasek (2015) with a sample of 170 convicted, exclusively adult male, sexual offenders. In their research, the variable of expeditious sexual history disclosure (supported by a truthful SHPE) was defined as achieving a truthful SHPE within 6 months of treatment onset. Bivariate analysis revealed that achieving a non-deceptive SHPE within 6 months of treatment onset was minimally correlated with sexual recidivism ($r\Phi$ = -.152, p = .047) as well as the variable of age of non-deceptive SHPE over/under age 35 ($r\Phi$ = .167 p = .029). Passing a SHPE (regardless of timeliness) was positively associated with the variable of treatment completion ($r\Phi$ = .328, p < .001). Again, the correlation among SHPE results and treatment completion was problematic because of the lack of independence between these variables.

Because these prior studies produced limited and mixed results -- primarily due to small sample sizes and methodological issues (e.g., defining expeditiousness categorically in 6-month and 12-month time intervals – and absence of variable independence relative to SHPEs and treatment outcomes) -- the researchers in the current study decided to change course. We combined samples to produce a larger sample size (N=280) consisting of exclusively adult male sexual offenders receiving community supervision, cognitive-behavioral sex offender treatment, and sexual history polygraph examinations. We also examined timeliness of passing a SHPE as a continuous/interval variable.

The importance of timeliness and veracity of client self-disclosure

Assessment of symptoms that relies upon honest and timely patient self-disclosure provides for intervention that may be the difference between speedy recovery and prolonged illness (Palmieri & Stern, 2009). In risk assessment of persons convicted of sexual offenses, the importance of verifying sexual history information, primarily through the extraction of data from official criminal records, has been well established (Hanson & Bussière, 1998). The impactful role of polygraph in eliciting what has been termed Clinically-Relevant Disclosures (CRDs) and Risk-Relevant Disclosures (RRDs) among sexual offenders has been acknowledged by scholars, and research indicates a substantial increase in CRDs and RRDs when polygraph is utilized in comparison to non-polygraph controls (Gannon, Wood, Pina, Barnoux & Vasquez, 2014; Wood, Alleyne, Ó Ciardha & Gannon, 2020).

Discerning the veracity of disclosed patient/client information is difficult and can be time consuming for evaluators, therapists, and supervision personnel. To illustrate, see the analysis provided by psychiatrists Palmieri and Stern (2009), concerning deception that occurs in the context of doctor-patient communication. These researchers indicate that deliberate acts of deceit, which include denying, distorting, obfuscating, fabricating, omitting, providing irrelevant information, and being non-responsive, damages the clinical relationship and compromises care. The consequences of unrecognized patient deception include failing to address key clinical issues, inappropriately responding to malingering, prescribing unneeded or harmful controlled

medication, and contributing to avoidance of deserved legal sanctions.

Farber and colleagues have extensively studied the process of patient self-disclosure in voluntary psychotherapy (Farber, 2003; Farber & Hall, 2002). In his comprehensive review of the disclosure literature, Farber found the following issues most prevalent among items not disclosed by clients: sexual and body-oriented experiences, sexual feelings, fantasies toward therapist, interest in pornography, bathroom habits, experiences and feelings toward masturbation, loss of virginity and fidelity (Farber, 2003). Farber questions the notion of full disclosure (a term that has been used in sex offender intervention for over 30 years), indicating that such disclosure is more of an ideal than an actuality. Farber prefers to use the terms extensive, salient and thorough, when describing the qualities of therapeutically valuable disclosure. The primary motivators for useful self-disclosure include therapeutic alliance, salience of topics discussed and the degree to which clients feel they are benefiting from therapy. Obstacles to timely and thorough self-disclosure include feelings of shame, guilt and fear (Farber, 2003; Farber & Hall, 2002). Many of the above dimensions of client disclosure apply to the assessment and treatment of individuals adjudicated for sexual crimes (Levenson, 2011; Marshall, Thornton, Marshall, Fernandez & Mann, 2001; Nunes, Hanson, Firestone, Moulden, Greenberg & Bradford, 2007).

The above information reveals that thorough, credible, self-disclosure of problematic health and behavioral conditions is vital in treatment planning and intervention – both with general mental health patients and with sexual offenders. Answering the question of how practitioners working with mandated clients determine if disclosure is timely enough, credible enough, and thorough enough to achieve successful treatment outcome and sexual recidivism is complicated and deserving of much more investigation. Sexual history polygraph testing is one strategy that may help provide intervention professionals with some degree of confidence that thorough, clinically relevant, disclosure has occurred.

Utilization of Polygraph in the Sexual History Disclosure Process

For sexual offenders proceeding through evaluation and treatment, thorough and relevant self-disclosure of offense details and deviant sexual history -- and ascertaining whether answers to relevant polygraph questions (which are constructed to target significant sexual history information the examinee is deliberately withholding) are deceptive or non-deceptive -- is a process that may take a short period of time (months) or a prolonged period (years). Aside from what a client may have disclosed during police, child services and attorney interviews, the initial assessment interview conducted by a sex offender treatment specialist, is the first opportunity for an individual to thoroughly discuss in detail his/her instant offense details and history of sexually deviant behavior. Other opportunities for disclosure during a client's evaluation and treatment trajectory, some of which involve polygraph testing, include:

a. the client filling out evaluation and treatment intake forms;

b. the individual completing a sexual history questionnaire;

c. the individual filling out paper/pencil or computerized psychometric tests;

d. the subject discussing his or her sexual history in group or individual therapy sessions;

e. the person conveying sexual history information during meetings with his or her parole/probation or other supervising officer;

f. the client providing information at treatment planning or risk assessment sessions with his or her therapist;

g. the examinee completing sexual history polygraph pre-test questionnaires;

h. the person verbally disclosing during the polygraph pre-test and post-test interviews with a polygraph examiner;

i. the individual debriefing during polygraph

post-test interviews with the treatment provider and/or probation/parole officer;

j. the person taking responsibility during community accountability, circles of support, or victim clarification sessions;

k. the subject providing necessary instant offense and sexual history information to sex offender registration and residence-monitoring authorities;

l. the individual disclosing during closing treatment progress assessments, post-treatment tests, aftercare planning sessions and treatment exit interviews.

Facilitating the disclosure opportunities listed above are various stakeholders in the offender's success (e.g., treatment providers, supervision officers and polygraph examiners) and their individual assessments of what disclosures are considered timely, thorough, relevant and credible. Although polygraph pretest and post-test interview processes may be powerful in elicitation of sexual history information, the elicitation efforts made by other stakeholders enumerated above are equally important. Westwood, Wood and Kemshall (2011) detail various techniques to elicit what they term responsible disclosures by sexual offenders (separate from what can be accomplished via polygraph testing). They acknowledge that during a time when mandatory polygraph testing is increasing, and resources are limited, disclosure elicitation techniques practiced by other stakeholders in the individual's success are important for positive outcomes.

Principles of practice for certified therapists working with sexual offenders provide some guidance relative to the utilization of polygraph to assist in elicitation of therapeutically relevant disclosures. The Association for the Treatment of Sexual Abusers (ATSA), an International organization comprised of approximately 3000 scholars and practitioners, representing approximately 20 countries, is concerned with making society safer through empirically supported evaluation, risk assessment, therapeutic and risk reduction/management practices with sex offenders (ATSA, 2020). The adult sex offender practice guidelines published by ATSA, last revised in 2014, indicate that the collection of sexual history information is an important component in clinical and case management practice (ATSA, 2014). Though cautious in its references to validity and reliability of sex offender polygraph testing, the authors of the ATSA practice guidelines acknowledge the following objectives that post-conviction polygraph testing (PCSOT) serves in the assessment, treatment, and risk reduction/management processes: 1. Facilitating a client's disclosure of sexual history information, which may include sexually abusive or offense-related behaviors (generally disclosed in the interview portion of the examination); 2. Eliciting from the client clarifying information regarding the instant/index offense; 3. Exploring potential changes, progress and/or compliance relative to treatment and other case management goals and objectives (through yes/no questions about adherence to specific treatment and other case management expectations); and/or 4. Making collaborative case management decisions about a client with other partners and stakeholders based on the information gleaned from the evaluation interview (ATSA, 2014, p. 75).

The above information suggests that sex offender polygraph testing can play an important role in the elicitation of important treatment and case management information. The time between the date of a sexual offender's initial disclosure to a treatment professional, and the date of passing a SHPE, could also be an indicator of intervention responsivity or perhaps the likelihood of sexual recidivism.

Research on the Accuracy of Specific Issue and Screening Polygraph Exams

Recent academic studies on the extent, nature and accuracy of information gained from polygraph testing are noteworthy (Gannon, Wood, Pina, Tyler, Baroux & Vasquez, 2014; (Wood, Alleyne, Ó Ciardha & Gannon, 2020). Gannon and colleagues (2014) revealed that the quantity of Clinically Relevant Disclosures (CRDs) improved significantly over self-disclosure sessions conducted without the polygraph – finding that polygraphed offenders made 572 CRDs in therapeutic sessions versus 320 CRDs for non-polygraphed controls. In a recent mixed methods study of supervisees, suspects and persons seeking relief from registration requirements (Wood, et al., 2020) found the following with respect to what they

termed Risk-Relevant Disclosures (RRDs):

• Voluntary or mandatory polygraphed supervisees were equally likely to make RRDs, but voluntary polygraph tests often failed to go ahead (*with re-testing after a deceptive, significant response [SR] polygraph result – added by these authors*).

• Relative to comparisons, supervisees undergoing polygraph testing (voluntary and mandatory) were nearly 6 times more likely to make at least one RRD.

• Supervisees across all levels of risk were more likely to make a RRD than comparisons.

• During polygraph sessions, polygraphed supervisees made more RRDs in the pre-polygraph interview than they did in the post-polygraph interview. Polygraph testresults revealing a significant response (i.e., indicative of an untruthful response) were associated with higher levels of post-polygraph interview RRDs.

• Polygraphed supervisees were more likely than comparisons to make RRDs regarding sexual interest in and/or increased access to children (online or offline). Comparisons were more likely to make RRDs regarding new relationships.

• In contrast with comparisons, polygraphed supervisees' RRDs resulted in morechanges to the focus of supervision (e.g., increase in home visits).

• Offender managers in the polygraph group rated the helpfulness of the polygraphas over 5 on a 7-point scale; regardless of whether RRDs had been made. The qualitative statements made by Offender Managers in interviews supported this. However, they were concerned about the voluntary nature of the polygraph resulting in test refusal (Wood, et al., 2020, p. *vii*).

Polygraph critics have questioned whether any benefits associated with pre-test and post-test utility of information gained via a disclosure process that incorporates polygraph testing outweigh the costs of relying on screening tests that are imperfect, not completely standardized, and/or not validated via peer-reviewed academic researchers (Ben-Shakhar, 2008; Iacono, 2008; Iacano & Ben-Shakhar, 2019; National Research Council, 2003; Rosky, 2012). These criticisms notwithstanding, some progress has been made in the past 10-15 years regarding these concerns. Studies have been published in Polygraph, now called *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* -- the peer-reviewed professional practice journal of the American Polygraph Association -- and in academic publications (Grubin, 2008, 2010; Honts & Thurber, 2019; Raskin & Kircher, 2014) that address the validity of forensic diagnostic polygraph tests and multiple issue screening exams.

Comprehensive reviews on the validity of CQT polygraph examinations (those most similar in design, administration and scoring to the CQT exams utilized in the current research project), have been published by several scholars. Honts (2004) found from four high quality field studies (*N*= 190) that overall accuracy of the CQT for testing specific issues is 90.5 per cent with nearly all errors (approximately 12%) being false positive errors. Ginton (2012) found that the single-issue CQT utilized in a field application produced decision accuracy of 94% for guilty and 84% for truthful subjects. Similarly, Raskin and Kircher (2014) an extensive review of CQT methods, scoring techniques, decision rules and classification accuracy, again provided evidence that specific issue evidentiary exams correctly classify approximately 92% of computer-scored cases. It is becoming clear that the CQT, when utilized properly by skilled examiners in forensic applications, can classify truthful and deceptive subjects with a relatively high degree of accuracy, and with relatively low false positive and false negative error rates (Raskin & Kircher, 2014).

Unfortunately, little published information exists on the validity of the CQT as applied in *sex offender screening* applications. The only research uncovered by these authors addressing the validity of CQT screening tests is the meta-analytic review of 14 studies involving 1,008 cases and 31 different polygraph scorers conducted by the American Polygraph Association

(2011). Specifically, American Polygraph Association (2011) researchers found that multiple issue CQTs scored with an assumption of independent criterion variance (tests similar to those used in post-conviction sex offender testing), produced a mean unweighted accuracy rate of .850 (85.0%), with a 95% confidence interval from .773 to .926, with an inconclusive rate of .125. Criterion accuracy values (deceptive decisions correct [87.3%], truthful decisions correct [83.1%], false positive errors [14.4%] and false negative errors [11.3%]) for such exams are detailed in (American Polygraph Association, 2011, p. 246).

The above research on polygraph accuracy and the utility of polygraph in eliciting disclosure that may improve treatment targeting and the management of sex offender risk shows advancement in the field, although much more work needs to be done in the area of sex offender screening. Gaps in knowledge remain relative to timeliness and thoroughness of polygraph-precipitated disclosure and its relationship to treatment completion and recidivism. This research project is intended to fill some of these gaps in knowledge.

Research Question, Variable Definition, Methods and Procedures

The primary question for this research project is whether timeliness of sexual history disclosure concomitant to passing a SHPE, is associated with sexual recidivism in this non-random sample. This is a correlation study – and to be clear – any correlational relationships found among variables, though informative, should not be generalized to larger populations or be interpreted as causal. To address the research question, these researchers conducted bivariate analyses on the independent study variables (Truthful SHPE, Timeliness of Truthful SHPE, and Treatment Completion) among the independent literature-derived variables (i.e., Static-99R risk score, presence of sexual deviance, denial, and psychopathy/ antisocial personality disorder) – as correlated with the dependent variable of *sexual recidivism.*

Study variables were defined as follows:

- *Truthful SHPE* – passing a sexual history polygraph examination (showing No Significant Responses [NSR] to all relevant sexual history polygraph test questions) as scored by a state-licensed, state corrections department approved, polygraph examiner experienced in testing sexual offenders. Coding: 1 = NSR SHPE ever achieved; 0 = NSR SHPE never achieved during the study period.

- *Timeliness of Truthful SHPE-* – months between the date of treatment program intake and the date of truthful SHPE. Coding: N/A – entered as the actual time value measured in months.

- *Treatment Status Completed* – completed treatment meeting all program requirements. Coding: 1 = completed treatment meeting all requirements; 0 = did not complete treatment during the study period (regardless of reason). Specifically, treatment completion included: fulfilling all treatment goals outlined in a treatment plan [including passing a SHPE and maintenance/compliance polygraph examinations], gaining maximum benefit from treatment [in the professional opinions of the program director, treating therapist, and treatment team – with input from supervising parole/probation officer] at time of discharge consideration, accomplishing treatment milestones relating to cognitive-behavioral components [such as arousal reconditioning], and complying with treatment rules/expectations). It should be noted that although passing a SHPE was one of many requirements for treatment completion, failing a SHPE was never the sole reason for unsuccessful discharge or non-completion of treatment.

- *Sexual Recidivism* – charged with or convicted of a sex crime documented in official criminal history records from three official sources (excluding status sexual offenses such as Failure to Register / Failure to Report offenses) occurring during the 5 years following program discharge. Coding: 1 = yes, sexual recidivism occurred within 5 years following program discharge; 0 = no, sexual recidivism did not occur within 5 years following program discharge.

Literature-derived control variables identified as associated with sexual recidivism were included in the analysis and defined as follows:

- *Risk Score* – score on Static-99R. Coding: N/A – collected as a continuous/interval

numeric value indicating the actual score on the Static-99R.

• *Risk Level* – categorical level of sexual recidivism risk as measured with the Static-99R (consistent with the categories defined in the Static-99R practice manual – see Harris, Phenix, Hanson, & Thornton, 2003). Coding: 1= low risk (scores of -3 to 1 on the Static-99R); 2 = moderate/low risk (scores of 2-3 on the Static-99R); 3 = moderate/high risk (scores of 4-5 on the Static-99R and 4 = high risk (scores of 6 and above on the Static-99R).

• *Denial* – documented failure (at initial program intake) to admit or assume responsibility for the commission of the instant offense (consistent with information contained in police reports, court documents, psychological or sexual deviance evaluation reports, presentence reports or treatment intake forms) and/or failure to admit documented allegations of other victims and other deviant sexual behaviors. It should be noted that none of the participants in this study took SHPEs if they were in categorical denial of their instant offense and they could not admit having at least one sexual victim. Coding: 1 = instant offense and/or sexually deviant behavioral history denial initially present; 0 = instant offense and/or sexually deviant behavioral history denial not present.

• *Sexual Deviance* – documentation in official records, evaluation reports, and presentence reports of deviant sexual arousal or interest (shown on a penile plethysmograph (PPG) and/or viewing time / deviant sexual interest examination (VT) report. Coding: 1 = the presence of sexually deviant arousal or interest; 0 = no sexually deviant arousal or interest documented.

• *Psychopathy or Antisocial Personality Disorder (APD)* -- because the variables of psychopathy and antisocial personality disorder (APD) were not delineated during data extraction from hard copy client files, one combined variable (Psychopathy or APD) was utilized. Coding: 1 = either present; 0 = neither present. Any interpretation of data emanating from the analysis of these erroneously combined constructs must be done cautiously.

Variable data were extracted from case files and program databases of two outpatient therapy programs specializing in the treatment of sexual offenders. Two research assistants were utilized to a) blind these researchers from recidivist identity, and b) alleviate any appearance of bias or conflict of interest (as recommended by O'Connell, 2011). State and national recidivism data was captured in 2011 and 2015 by the research assistants from three official sources (i.e., Oregon Judicial Department, Justice Information Network [OJIN], 2015; LexisNexis Accurint, 2015; Washington State Patrol, 2011 and 2015).

The data extracted from archived polygraph examination hard-copy reports and pretest questionnaires were then entered into the study database by the research assistants. Static-99R (Hanson and Thornton, 2000; Harris, Phenix, Hanson and Thornton, 2003; Helmus, Babchishin, Hanson and Thornton, 2009) risk computation forms were completed by the lead investigator of this study after he received official training in 2011 on proper scoring of the Static-99R from the Justice Institute of British Columbia (JIBC, 2011). These data were entered into the study database prior to the research assistants matching recidivism data to the identified dataset.

Specific data on the quantity and quality of sexual history disclosures (e.g., number of previously undisclosed victims, number and seriousness of unknown sexual offenses, and the number and nature of previously undisclosed paraphilias) was extracted on the 112 case files added to this study from the original dataset (n=170) examined in 2015 (Konopasek & Nelson, 2015) and are provided below. Unfortunately, quantity and quality of disclosure was not collected on the 2015 dataset, so aggregation of that data is not possible.

Assumptions

It is assumed that the polygraph screening examinations conducted on individuals in this sample (all of which were CQT exams) are similar in design, administration, and criterion accuracy to those described by (Honts, 2004, 2017; National Research Council, 2003; Amer-

ican Polygraph Association, 2011; Raskin & Kircher, 2014). The polygraph examination reports reviewed by the researchers in the current study revealed that relevant questions contained in the sexual history polygraph examinations were generally constructed as follows:

- As an adult, have you engaged in sexual contact with any child you have not disclosed?

- Besides what you told me, have you engaged in sexual contact with any person against his/her will?

- Have you committed a sex crime (contact or non-contact) on any person you are keeping secret?

The above examples of relevant questions are constructed to take into account, and exclude, sexual history disclosures made by examinees prior to the SHPE (generally during the polygraph pretest interview) and are expected to be answered "No" by the examinee. A deceptive call or finding is made when significant physiological responses by the examinee are produced on the polygraph exam to any such relevant questions -- termed "SR" by the polygraph examiner. A non-deceptive call is made by the examiner when an examinee's "No" answers to all such relevant questions produces no significant physiological responses on the polygraph exam – termed "NSR" by the polygraph examiner.

It should be noted that because a substantial portion of the polygraph examinations administered on this sample occurred prior to publication and implementation of the American Polygraph Association's policies and standards for Post-Conviction Sex Offender Testing (PCSOT), the exams did not separate what is now termed in polygraph practice as the Sex History I portion of a SHPE (undisclosed sexual offenses) from the Sex History II portion of a SHPE (problem sexual behaviors, deviant sexual acts, and paraphilias). The examinations conducted with the relevant questions delineated above therefore focused more on Sexual History I issues (see American Polygraph Association, 2018, for information).

To reiterate, the CQT sexual history polygraph examinations utilized on this study population were similar in structure to screening tests validated by American Polygraph Association (2011) researchers. All exams were conducted by experienced, state-licensed and/or state department of corrections approved, independent polygraph examiners – and it is assumed that all followed best practices for testing sexual offenders in existence at the time of the study. None of the examinations on subjects included in this study were conducted by the lead author of this study, who is now a state-licensed and federal polygraph examiner.

Sample Characteristics and Descriptive Statistics

The study population is a convenience sample of all adult male sexual offenders (N=280) referred to two outpatient treatment programs (utilizing cognitive-behavioral treatment techniques in accordance with ATSA standards) – programs that were directed/administered by the authors of this study in separate locations. All participants completed at least one sexual history polygraph examination (SHPE) as part of supervision mandates and treatment requirements. Participants either completed evaluation and/or treatment before year-end 2009 or were terminated from treatment before year-end 2009.

The mean age of participants taking SHPEs was 35.96 years (range of 18 to 80 years) and the mean number of months needed to achieve a passed SHPE was 6.41 months (S.D. 10.83 months, range 0 to 91 months). Of the 202 participants who achieved a truthful SHPE, a substantial number passed a SHPE on their first polygraph test-taking attempt (n = 147), 52.5% of the total sample. Forty-three (43) individuals (15.4% of the total sample) passed on a 2nd polygraph test-taking attempt; seven (7) individuals (2.5% of the total sample) on their 3rd attempt; and three (3) participants (1.1% of the total sample) on a 4th polygraph test-taking attempt.

Although test/retest habituation on the individuals who were administered more than one polygraph examination was a concern, the decision was made to keep these individuals in the research sample for the following reasons: every CQT examination is different relative to the pre-test interview and comparison ques-

tions, not all persons re-taking a SHPE were tested by the same examiner, and retests were generally not administered without additional post-test disclosure being made by examinees. Unfortunately, data on any disclosures elicited from multiple test-taking attempts was not separated from the total number of disclosures made prior to finally passing a SHPE. Comparative data on the deceptive (SR) and non-deceptive (NSR) groups is indicated in Table 1 below.

Table 1 Group Characteristics of Deceptive (SR) and Non-deceptive (NSR) SHPE Participants

Characteristic	Deceptive (SR) (*n*=78)	Non-Deceptive (NSR) (*n*=202)
Mean Age at time of SHPE	38.40 years	35.02 years
Sexual Recidivism	10 (12.82%)	12 (5.9%)
Mean Static 99-R Risk Score and (Level) High Risk Level on Static-99R	2.23 (Moderate-Low) 6 (7.7%)	2.32 (Moderate-Low) 9 (4.5%)
Mean Timeliness of NSR SHPE in months	----	6.41
Presence of Sexual Deviance via Documented PPG or VT Testing	43 (55.12%)	82 (40.59%)
Denial Present (Initially)	67 (85.90%)	73 (36.14%)
Presence of Documented APD / Psychopathy	12 (15.38%)	29 (14.36%)

Table 1 reveals that individuals who never passed a SHPE during the treatment period of this study had a substantially higher sexual recidivism rate (12.82%) than non-deceptive participants (5.9%). Participants who produced deceptive SHPE results also had a substantially higher incidence of initial denial than non-deceptive subjects (85.9% versus 36.14%), which might suggest fear of making disclosures or admissions, recalcitrance, or other barriers to making responsible disclosures and admissions. As for risk level, SR (deceptive) and NSR (non-deceptive) subjects produced very similar mean risk scores on the Static-99R, falling into the Moderate-Low risk category; however, a higher percentage of SR deceptive individuals (7.7% versus 4.5%) fell into the high risk category.

Table 2 displays the data collected from the added dataset (n=112) containing disclosure information on both deceptive and non-deceptive individuals.

Table 2 Additional Disclosures of Deceptive (SR) and Non-deceptive (NSR) SHPE Participants (n=112)

	Deceptive – SR (n=24)				Non-deceptive – NSR (n=88)			
Disclosure Type	Sum	Mean	Medn	[Min-Max]	Sum	Mean	Medn	[Min-Max]
Male Child Victims	13	.54	0.0	[0-4]	73	.83	0.0	[0-27]
Female Child Victims	263	10.96	4.0	[0-118]	1312	14.91	4.0	[0-433]
Adult Rape Victims	10	.42	0.0	[0-3]	75	.85	0.0	[0-10]
Other Contact Victims	177	7.37	0.0	[0-100]	175	1.99	0.0	[0-100]
Non-Contact Victims	19	.79	0.0	[0-6]	880	10.00	0.0	[0-500]
Paraphilic Acts	51	2.13	0.0	[0-16]	3230	36.70	1.0	[0-400]
Child Porn Image/Vid	328	13.67	0.0	[0-215]	2027	23.03	0.0	[0-1000]

With respect to additional pretest disclosures (above and beyond the instant offense), there are substantial differences between deceptive and non-deceptive individuals. Non-deceptive individuals produced substantially more total disclosures and a higher mean number of additional disclosures in every category except "Other Contact Victims" – which in the vast majority cases involved acts of non-consensual sexual touching of adults or unknown age victims. There is also evidence of individual cases, considered outliers (as shown by some extraordinarily high values in the Max column) – attributed to persons perpetrating acts of frottage or indecent exposure to numerous adult and child victims – skewed some of the mean statistics shown in Table 2.

Individuals completing treatment comprised 50.7% (*n* = 140) of the study sample, whereas 49.3% (*n* = 138) were discharged by therapists. Two participants deceased after completing their respective treatment programs and during the recidivism follow-up period. Except for recidivism data generated on the individuals in this sample who did not complete treatment during the treatment time frame of this study, little is known about their eventual treatment outcome or whether these individuals ever passed a SHPE.

Findings

The criminal history record checks conducted in late 2014 and early 2015 revealed a wide array of offenses perpetrated by recidivists, ranging from Indecent Exposure, to Child Pornography, to Child Sexual Abuse/Child Molestation, to Forcible Rape, and are shown in Table 3 below. Again, the variable of sexual recidivism was defined in this study as having been charged with, or convicted of, a new sexual offense (excluding status offenses such as failure to register/report), documented within official criminal history records, within the 5-year recidivism follow-up period following program discharge. Table 2 outlines the most serious sexual offenses perpetrated by known recidivists.

Table 3 Most Serious Recidivism Offenses (During 5-Year Post Intervention Follow-Up Period)

Sexual Recidivism Offense	Frequency
Child Molestation / Sexual Abuse of Child	*n* = 9
Child Pornography Offenses	*n* = 5
Forcible Rape	*n* = 5
Rape of a Child	*n* = 1
Prostitution	*n* = 1
Public Indecency	*n* = 1
Total 5-Year Sexual Recidivists	*n* = 22 (7.9% of *N* = 280)

Although the known sexual recidivism rate for the entire sample was relatively low (7.9%), a substantial number of the 22 subjects recidivating during the follow-up period were convicted of crimes against children (*n*=16, 72.7%).

Bivariate correlational analyses were conducted utilizing the Phi-Coefficient for dichotomous categorical study variables and any association with the variables of treatment completion and sexual recidivism. Pearson Correlation was used to evaluate the relationship between the continuous variables (Timeliness of Truthful SHPE and Static-99R risk score) as associated with treatment completion and sexual recidivism. Bivariate correlations are shown in Table 4 below.

Table 4 Bivariate Correlations Among Study Variables and Literature Derived Control Variables with Sexual Recidivism and Treatment Completion

	Value	Sig.	Value	Sig.
	Sexual Recidivism		Treatment Completion	
Truthful SHPE (ever - regardless of timeliness, *n* = 202)	-.115	.055	.328**	.000
Timely Truthful SHPE	.017	.776	.155**	.009
Static 99-R Risk Score	.167**	.005	N/A	N/A
Presence of Sexual Deviance	.192**	.001	-.207**	.000
Treatment Outcome (Completed)	-.137*	.022	-----	----
Presence of Denial	.053	.376	-.114	.056
Presence of Psychopathy / APD	-.008	.890	-.036	.546
Total Cases				280

*Significant at $p < .05$ (2-tailed)
**Significant at $p < .01$ (2-tailed)

Table 4 indicates that the polygraph-related study variables (Truthful SHPE and Timely Truthful SHPE) are statistically significant and correlated with treatment completion in the expected positive direction. This is not surprising given the fact that passing a SHPE within 6 to 12 months of program entry was valued as an important milestone in the treatment programs from which this sample originates. We can therefore not assume that these variables are independent of one another. Further, Table 4 reveals that the variable of Timely Truthful SHPE (treated as continuous variable) was not associated with sexual recidivism (r = .017, p = .776). The analysis also reveals that the correlation between the variables of treatment completion and sexual recidivism are statistically significant and correlated in the expected negative direction ($r\Phi$ = -.137, p = .022). As for the variables of Static-99R score and sexual deviance, both are correlated with sexual recidivism, consistent with what has been shown in prior research. In this sample, no statistically significant correlations were found among the literature derived variables of denial and the presence of Psychopathy/APD, with sexual recidivism.

To determine whether the statistically significant independent variables (treatment completion, Static-99R risk score and presence of sexual deviance) could be useful in a model that predicts sexual recidivism, these variables were analyzed via logistic regression. Logistic regression is the best statistical analysis when the dependent variable is dichotomous/binary and predictor/independent variables are of differing levels of measurement such as interval/ratio or nominal/dichotomous (Hosmer and Lemeshow, 2000; Tabachnick and Fidell, 2007; UCLA Academic Technology Services, 2009). Results of the logistic regression analysis are indicated in Table 5 below.

Table 5 Logistic Regression – Statistically Significant Variables Correlated with Sexual Recidivism

Variables	*B*	S.E.	Wald	df	Sig	*Exp(B)*	95% CI *Exp(B)* Lower	Upper
Treatment Completed	.703	.516	1.857	1	.173	2.020	.735	5.554
Static-99R Score	.243	.121	4.074	1	.044	1.276*	1.007	1.616
Sexual Deviance	-1.294	.537	5.80	1	.016	.274*	.096	.786
Constant	-3.051	.576	28.030	1	.000	.047		

Significant at $p < .05$

Table 5 shows that the variables of Static-99R and sexual deviance are statistically significant within the logistic regression model; however, the variable of treatment completion loses significance as a possible predictor within the model (*Exp[B]* = 2.020, p = .173). The odds ratios for the relationships between the independent variables of Static-99R score and Sexual Deviance with the variable sexual recidivism are (*Exp[B]* = 1.276, p = .044) and (*Exp[B]* = .274, p = .016) respectively. These findings can be interpreted as follows: a) for every one unit increase in Static-99R score, the odds of sexual recidivism increases by a factor of .276; and b) for every one unit increase in the measure of sexual deviance, the odds of falling into the sexual recidivist category is increased by a factor of .274. Relative to case classification, this logistic regression analysis correctly predicted 16 of the 22 sexual recidivists (72.7%) and 175 of 258 (67.8%) of non-recidivists, with an overall accuracy of 68.2%.

Discussion and Limitations

Regarding any relationship between the timeliness of passing a SHPE (analyzed as an interval/scale variable of time in months between treatment onset and date of truthful SHPE) and sexual recidivism, our results show that the variables are not statistically significant or correlated. This differs from earlier research findings that show statistically significant associations among the variables when the variable of timely SHPE is categorically defined in intervals of 6 and 12 months (Konopasek, 2015; Konopasek & Nelson, 2015). Variables that evaluate the timeliness of truthful sexual history polygraph examinations are certainly, in the view of these authors, worthy of further inquiry. Our bivariate and logistic regression analyses produced findings that are consistent with the work of several researchers in the sex offender risk assessment field (Hanson & Bussiere, 1998; Hanson & Thornton, 2000; Harris, Phenix, Hanson & Thornton, 2003; Helmus, Babchishin, Hanson & Thornton, 2009), showing that Static-99R score and presence of sexual deviance are statistically significant predictors of sexual recidivism.

Our findings also reveal that the variable of treatment completion (defined as fulfilling all treatment goals outlined in a treatment plan [including passing a SHPE], gaining maximum benefit from treatment [in the professional opinions of the program director, treating therapist and treatment team] at time of discharge consideration, accomplishing treatment milestones relating to cognitive-behavioral components, and complying with treatment rules/expectations) -- though correlated with the variable of sexual recidivism in the bivariate analysis, lost statistical significance within the logistic regression model. It may be that any bivariate correlation between treatment outcome and sexual recidivism is so weak that any correlational influence is lost when treatment completion is included in an analysis of other variables (e.g., Static-99R and measures of sexual deviance) that have greater effects on sexual recidivism. Further, because the current study added cases to the dataset utilized by (Konopasek & Nelson, 2015), it is not surprising the current results (garnered from bivariate analysis) are consistent with the earlier findings -- that the variables of passing a SHPE, and timeliness of passing a SHPE, are associated with treatment completion. As noted earlier, findings in this study relative to any relationship between SHPE performance and treatment outcome are tenuous at best -- because of the interrelatedness of these variables within this combined sample.

This study has several limitations that necessitate cautious interpretation of findings. First and foremost, this is a correlational study in which no non-polygraphed control group was utilized; therefore, these findings should not be generalized to larger groups and should not be interpreted as causal. Another limitation has to do with the fact that a convenience sample of sexual offenders (N=280), containing sexual offenders who were available and included in the study because all were admitted to their respective treatment programs, and all took at least one SHPE, makes the sample subject to selection bias. Related to this is that the treatment programs were administered/directed by the authors, who are therefore non-independent from decisions about completion of treatment. Random sampling of much larger client populations (treated and untreated, polygraphed and not polygraphed) – and statistically comparing the samples with one another -- would have been a much better methodological approach. As with any sexual recidivism study, samples of persons convicted of committing offenses are limited by the numbers of recidivating individuals who are available for study (often referred to as the base rate dilemma). Even when every case participating in treatment and SHPEs is included in the study (as was the case with this sample) obtaining enough recidivist participants to draw meaningful conclusions is difficult. In this study, all 280 adult males who were mandated to complete SHPEs and complete sex offender treatment were included in the study and yet the sample produced only 22 sexual recidivists during the 5-year post discharge follow-up period. This study would have been much more robust had we had the opportunity to analyze a larger sample that would have yielded more recidivists and provided for greater statistical power.

Another limitation has to do with utilizing several bivariate analyses to narrow the field of variables (i.e., those shown to be statistically significant relative to any correlation with sex-

ual recidivism) to provide enough statistical power to conduct LR analyses on three or four independent variables and one dichotomous dependent variable. This procedure excluded both SHPE variables, the variable of denial, and the erroneously constructed variable of (Psychopathy/APD). We acknowledge that separate and distinct measures of Psychopathy and APD might have yielded another statistically significant variable or two to include in the LR analysis. We also acknowledge that re-defining the variable of timely truthful SHPE as an interval/scale variable is more methodologically sound than utilizing categorical measurements of timeliness as done in previous research (Konopasek, 2015; and Konopasek & Nelson, 2015). It was certainly not the intentions of these researchers to give the appearance of attempting to data dredge or otherwise manipulate or influence findings (in this case to produce a statically non-significant SHPE finding relative to sexual recidivism).

A final limitation to this study is that the treatment outcomes of clients who discontinued treatment within the period of this study -- and whether those individuals eventually passed a SHPE or completed treatment at later dates -- are unknown. The ability to longitudinally track individuals who discontinued one of the subject treatment programs would likely to have made a difference in these findings.

To conclude, a great deal of caution should be exercised to not interpret these correlational findings from a convenience sample as generalizable to larger populations of sexual offenders, or as suggestive of any causal relationships among variables. At the same time, our findings should not be interpreted as evidence that analyzing SHPE-related variables (especially those that integrate measures of polygraph-precipitated sexual history disclosure and timeliness of passing a SHPE) are futile and should be discarded as variables for future inquiry. Rather, our project should be viewed as providing some limited knowledge on how to define and quantify SHPE-related variables, ways to test (and not test) such variables, and perhaps as a source of learning that can improve SHPE and recidivism related research methodologies in the future.

Recommendations for Future Research

Based on the above information, our recommendations for future work include the following:

1. Conduct SHPE-related sexual recidivism studies with larger samples that employ random sampling, and evaluate polygraphed and non-polygraphed control groups;

2. Separate and analyze PCSOT Sex History I disclosures and Sex History II disclosures relative to treatment completion and sexual recidivism. Include measures of sexual history disclosure obtained via multiple test-taking attempts -- disclosure that may occur between the time of initial deceptive SHPE and an eventual non-deceptive SHPE;

3. Further examine how timeliness of passing a SHPE may relate to the odds of a sexual offender falling into the sexual recidivist classification category;

4. Analyze of the timeliness of passing a SHPE with respect to motivating sexual offenders to make clinically relevant disclosures (CRDs) and risk-relevant disclosures (RRDs) -- utilizing variable definitions consistent with current Post-Conviction Sex Offender Testing [PCSOT] Sex History I and Sex History II classifications;

5. Conduct further research on the current study sample population, at a 10-year recidivism follow-up period, which will likely produce a larger base rate of recidivists, to determine whether the correlates identified as statistically significant (and not statistically significant) in the current project replicate over time. Perhaps include the dataset utilized in the current study in a meta-analytic review.

References

Abrams, S., & Ogard, E. (1986). Polygraph surveillance of probationers. Polygraph, 15(3), 175-182.

American Polygraph Association. (2018). Model Policy for Post Conviction Sex Offender Testing (PCSOT). Retrieved from: https://www.polygraph.org/assets/docs/Misc.Docs/PCSOT_Model_Policy_March_2018%20.pdf

American Polygraph Association, Ad-hoc Committee on Validated Techniques (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. Polygraph 40 (4).

Association for the Treatment of Sexual Abusers - ATSA. (2020). History of ATSA. Retrieved from: https://www.atsa.com/meet-atsa/history

Association for the Treatment of Sexual Abusers-ATSA. (2014). *Practice standards and guidelines for the evaluation, treatment and management of adult male sexual abusers.* Retrieved from http://www.atsa.com/Public/Adult/2014_ATSA_Adult_Guidelines_TOC.pdf .

Ben-Shakhar, G. (2008). The case against the use of polygraph examinations to monitor post-conviction sex offenders. *Legal & Criminological Psychology, 13*(2), 191-207. doi:10.1348/135532508X298577.

Cook, R. (2011). Predicting recidivism of the convicted sexual offender using the polygraph and Static-99. (Doctoral Dissertation). Retrieved from UMI Dissertation Publishing, ProQuest, (UMI Number: 3445252).

Cook, R., Barkley, W., & Anderson, P. B. (2013). The Sexual History Polygraph Examination and Its Influences on Recidivism. Journal of Social Change, 5(1), 1.

Farber, B.A. (2003). Patient self-disclosure: A review of the research. *Journal of Clinical Psychology, 59* (5), 589-600.

Farber, B.A. & Hall, D. (2002). Disclosure to therapists: What is and is not discussed in psychotherapy. *Journal of Clinical Psychology, 58* (4), p 359-370.

Gannon, T.A., Wood, J.L., Pina, A., Tyler, N., Barnoux, M.F.L., & Vasquez, E.A. (2014). An evaluation of mandatory polygraph testing for sexual offenders in the United Kingdom. *Sexual Abuse: A Journal of Research and Treatment, 26* (2) 178-203.

Ginton, A. (2012). A non-standard method for estimating accuracy of lie detection techniques demonstrated on a self-validating set of field polygraph examinations. *Psychology Crime and Law.* (19), 577–594. doi:10.1080/1068316x.2012.656118

Ginton A. (2017). Examining different types of comparison questions in a field study of CQT polygraph technique: Theoretical and practical implications. *Journal of Investigative Psychology and Offender Profiling.* (14), 281–293. https://doi.org/10.1002/jip.1475

Grubin, D. (2008). The case for polygraph testing sex offenders. *Legal & Criminological Psychology, 13*, 177-189. doi:10.1348/135532508X295165

Grubin, D. (2010). The polygraph and forensic psychiatry. *Journal of American Academic Psychiatry Law, 38*, 446–451.

Handler, M., Nelson, R., & Blalock, B. (2008). A focused polygraph technique for PCSOT and law enforcement screening programs. *Polygraph, 37*(2), 100-111.

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*(2), 348-362. doi:10.1037/0022-006X.66.2.348

Hanson, R.K., & Morton-Bourgon, K. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*(6), 1154-1163. doi:10.1037/0022-006X.73.6.1154

Hanson, R. K., & Thornton, D. (2000). Static 99: *Improving actuarial risk assessments for sex offenders.* Ottawa: Department of the Solicitor General of Canada.

Harris, A., Phenix, A., Hanson, R. K. & Thornton, D. (2003). Static-99 Coding Rules: Revised 2003, the Static-99R. Ottawa, ON: Solicitor General of Canada. Printed by the Justice Institute of British Columbia, Corrections and Community Justice Division.

Heil, P., Ahlmeyer, S. & Simons, D. (2003). Crossover sexual offenses. *Sexual Abuse: A Journal of Research and Treatment, 15*(4).

Helmus, L., Babchishin, K. M., Hanson, R. K. & Thornton, D. (2009). Static-99R: Revised age weights. Retrieved January 10, 2011 from: www.static99.org

Hindman, J., & Peters, J. (2001). Polygraph testing leads to better understanding adult and juvenile sex offenders. *Federal Probation, 65*(3), N. PAG.

Honts, C. R. (2004). The psychophysiological detection of deception. In P. A. Granhag, & L. A. Stromwall (Eds.), *The detection of deception in forensic contexts* (pp. 103-123). Cambridge, England: Cambridge University Press.

Honts, C. R., &Thurber, S. (2019, March). A Comprehensive Meta-analysis of the Comparison Question Polygraph Test. Paper presented at the annual meeting of American Psychology Law Society. Portland, Oregon.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Iacono, W. G. (2008). Accuracy of polygraph techniques: Problems using confessions to determine ground truth. *Physiology & Behavior, 95*(1-2), 24-26. doi: DOI: 10.1016/j.physbeh.2008.06.001

Justice Institute of British Columbia – JIBC. (2011). Static-99R: Sex Offender Risk Assessment. Retrieved January 10, 2011 from http://www.jibc.ca/course/soap105

Konopasek, J.E. (2011). Micro-level social learning correlates of sex offender recidivism: Expeditious sexual history disclosure via polygraph testing. (Doctoral Dissertation), Capella University. ProQuest, UMI Dissertations Publishing. Retrieved from: http://search.proquest.com/docview/908602302

Konopasek, J.E. (2015). Expeditious disclosure of sexual history via polygraph testing: Treatment outcome and sex offense recidivism. *Journal of Offender Rehabilitation, 54*(3), 194-211. doi:10.1080/10509674.2015.1023481

Konopasek, J. E. and Nelson, R. (2015). Sexual History Disclosure and Sex Offender Recidivism. *Polygraph, 44*(2). Retrieved from: https://www.polygraph.org/assets/docs/APA-Journal.Articles/2015/442%20konopasek%20and%20nelson%202015%20renumbered.pdf

Krapohl, D. (2006). Validated polygraph techniques. *Polygraph, 35*(6), 149.

Levenson, J.S. (2011). "But I didn't do it!": Ethical treatment of sex offenders in denial.

Sexual Abuse: A Journal of Research and Treatment, 23 (3), 346-364.

LexisNexis Accurint (2015). Public record criminal history records search program description. Retrieved May 9, 2011 from: http://www.accurint.com.

McGrath, R. J., Cumming, G., Hoke, S. E., & Bonn-Miller. (2007). Outcomes in a community sex offender treatment program: A comparison between polygraphed and matched non-polygraphed offenders. Sexual Abuse: A Journal of Research and Treatment, 19(4), 381.

Marshall, W.L., Thornton, D., Marshall, L.E., Fernandez, Y., & Mann, R. (2001). Treatment of sexual offenders who are in categorical denial: A pilot project. *Sexual Abuse: A Journal of Research & Treatment, 13*, 205-215.

National Research Council of the National Academies of Science, Committee to Review the Scientific Evidence on Polygraph (Ed.). (2003). The polygraph and lie detection (D.O. Sciences ed.). Washington DC: The National Academies Press.

Nunes, K. L., Hanson, K. R., Firestone, P., Moulden, H. M., Greenberg, D. M., & Bradford, J. M. (2007). Denial predicts recidivism for some sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 19*(2), 91.

O'Connell, M.A. (2011). Recommendations for research methodology to avoid conflict of interest, dual relationships and vested interest: The utilization of independent third party research assistants. Personal communication on February 26, 2011.

Oregon Judicial Information Network / Oregon Case Information Network (2015). Oregon judicial information network [OJIN], access to Oregon cases on the internet. Retrieved January 08, 2017 from: http://courts.oregon.gov/OJD/OnlineServices/OJIN/Pages/index.aspx

Palmieri, J.J. and Stern, T.A. (2009). Lies in the Doctor-Patient Relationship. *Primary Care Companion, Journal of Clinical Psychiatry.*

Proulx, J., Pellerin, B., Paradis, Y., McKibben, A., Aubut, J., & Ouimet, M. (1997). Static and dynamic predictors of recidivism in sexual aggressors. S*exual Abuse: A Journal of Research and Treatment, 14*(3), 257-284.

Raskin, D. C. & Kircher, J. C. (2014). Validity of Polygraph Techniques and Decision Methods. D. C. Raskin, C. R. Honts & J. C. Kircher (Eds.). *In Credibility Assessment Scientific Research and Applications* (First Edition), San Diego, CA: Academic Press - Elselvier

Rosky, J.W. (2012). The (f)utility of post-conviction polygraph testing. Sexual Abuse: A Journal of Research and Treatment, 25(3), 259-281.

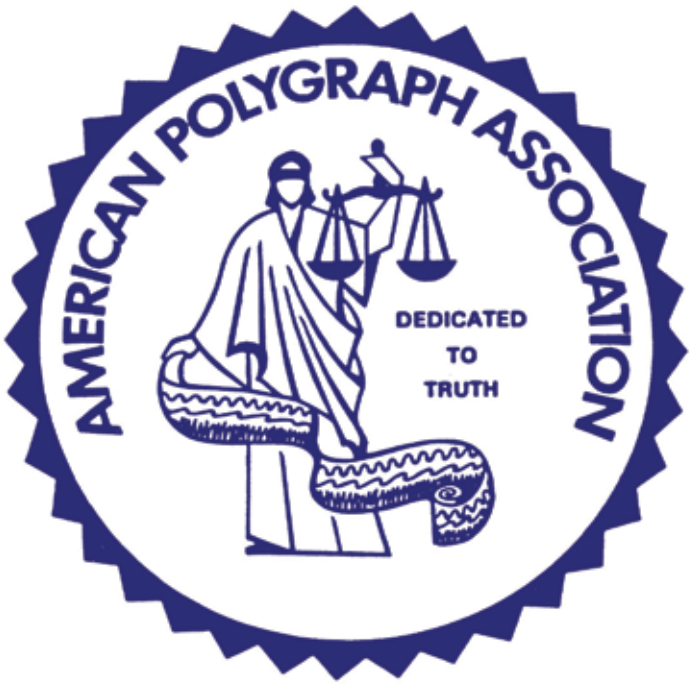
Tabachnick, B. C., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

UCLA Academic Technology Services. (2009). *SPSS data analysis examples. logistical regression.* Retrieved June 11, 2009 from http://www.ats.ucla.edu/stat/spss/dae/logit.htm

Washington State Patrol (2011 and 2015), Washington Access to Criminal History – WATCH, Retrieved May 1, 2011 from: https://fortress.wa.gov/wsp/watch

Westwood, S., Wood, J., & Kemshall, H. (2011). Good practice in eliciting disclosures from sex offenders. *Journal of Sexual Aggression, 17*(2), 215-227.

Wood, J. L., Alleyne, E., Ó Ciardha, C., Gannon, T. A. (2020) An Evaluation of Polygraph Testing by Police to Manage Individuals Convicted or Suspected of Sexual Offending. University of Kent (KAR id:81207)


AMERICAN POLYGRAPH ASSOCIATION
DEDICATED
TO
TRUTH