A Field Assessment of Automated Presentations of Polygraph Test

Donald J. Krapohl[1]

Donnie W. Dutton[2]

Dani Pruett[3]

Abstract

Characteristics of polygraph examinations from a large polygraph program were coded to help determine whether there were any effects on polygraph decisions that might be attributable to the use of automation to present test questions during the testing phase of a polygraph examination. Among the 415 cases in this six-month exhaustive sample, a small effect was found for the number of test charts the examiners recorded between the automated and human conditions. No significant differences were found among the proportions of polygraph decisions when comparing examinations in which the examiner read the test questions to examinations in which the computer presented them. The study found no adverse effects for the use of the digital voice in testing within the constraints of the variables tested. Given the advantages that automated presentation of test questions offers for standardization and a more useful allocation of examiner attention, its use in field polygraph examinations warrants consideration.

[1]Director for Educational Services, Capital Center for Credibility Assessment, and regular contributor to this journal.

[2]Mr. Dutton is an APA Past President and the Vice President of the Capital Center for Credibility Assessment.

[3]Ms Pruett is one of the first four UK police polygraph examiners. Now employed by Behavioural Measures UK (BMUK).

The authors appreciate the helpful comments and suggestions provided by the editor and the anonymous reviewer to an earlier version of this paper.

This article is one in a series titled Best Practices. The views expressed are solely those of the authors and do not necessarily represent those of their employers or the American Polygraph Association. Comments can be sent to the first author at APAkrapohl@gmail.com.

Since their introduction more than 30 years ago, digital polygraphs have come to replace pen-and-ink analog instruments that were the mainstay for most of the history of the polygraph. As early as 1962 Dr. Joseph Kubis[4] was conducting feasibility studies on computerization in polygraph exams for the US Air Force, and the first computer-assisted polygraph was developed in the late 1980s (Raskin & Kircher, 1990). New capabilities brought about by the transition to computer polygraphs are many: decision-support algorithms, post hoc data processing, feature extraction, electronic file sharing, and automating portions of the examination process. It is this last capability that is the focus of the present project.

All the major suppliers of computer polygraphs have software capable of presenting the test questions during testing using an automated voice. Replacing the examiner's voice with the automated voice for presentation of test questions has certain advantages. Automated voices are consistent across all questions, alleviating a concern regarding unconscious emphasis on certain test questions or accusations of such emphasis. Timestamping of question onsets is more reliable with automated question presentations because the software initiates the question presentation precisely at the point the examiner presses the key to signal question onset. The reliability of timestamping of question onsets based on the human voice requires coordination between an examiner's voice and the examiner's ability to simultaneously initiate the event marker, a skill that is likely to vary among different practitioners. The use of automated timestamping with the digital voice offers more confidence in the displayed latencies between question onset and response onset, a factor that is considered when assessing whether a response is associated with a test question. Also, allowing the computer to present the test question reduces attention demands on examiners, thus freeing them to visually monitor examinees more closely, perhaps affording more opportunities to detect countermeasures. For programs that have many polygraph examiners, the automated voice reduces variability among examiners, providing a more uniform testing experience for all examinees. Automated voices are also easily presented through headphones, with the additional advantage of removing extraneous sounds that may induce physiological responses. Finally, there is tentative evidence that automated presentation of test questions may improve polygraph decision accuracy (Honts & Amato, 2007).

There are unknowns regarding whether there are differential effects between human and automated question presentations. It could be possible examinees process questions differently if they are asked by a human compared to the same questions being presented by a machine. If, for example, examinees are more comfortable lying to a computer than they are lying to a human, or the reverse, a differential effect may appear in test outcomes. We found no reports that speak to whether the automated option has an adverse effect on the proportion of inconclusive polygraph results, or the number of test charts necessary to avoid inconclusive results. Similarly, the literature is silent as to whether a digitized voice corresponds with higher rates of decisions of deceptiveness or decisions of truthfulness.

In this six-month project we recorded polygraph test outcomes for examiners who used the digital voice and another sample of examiners from the same organizations who used their own voices during testing. While ground truth was not available for most of these cases, we did search for differential effects on test results and the number of test charts.

Method

Testing Examiners

There were 47 examiners in this project. All were trained by the same polygraph education

[4] Perhaps of historical interest, Professor Kubis assumed Father Walter Summers' position at Fordham University when the latter prematurely died in 1938. Reverend Summers, as all students of polygraph history will recognize, was a researcher known for conducting deception tests using a recording electrodermal device, and for his use of "emotional standards," technical questions that approximate what today would be called "comparison questions."

program and participated in the same quality control oversight program. Among the 47 testing examiners, 25 used digitized voice and 22 presented the questions with their own voices. Examiners were free to choose which voice to use during their examinations. The examiners who used digitized voice (DV) had an average of 2.7 years of experience at the beginning of this project, with a range of two months to eight years. Examiners using their own human voices (HV) to present test questions had an average of 4.0 years of experience, with a same range as those who used digitized voices. The difference in experience between the groups fell short of statistical significance, $t(45) = 1.90$, $p = .06$, ns.

Cases

The period of data collection was from January 1 through June 30, 2022. A total of 415 cases were submitted for quality control review from a large offender management program and all were included in this study. The polygraph results decided by the quality control reviewers were used in place of those of the testing examiner if there were differences in results between them. This occurred in 44 cases, or 10.5% of the sample. The Empirical Scoring System provided the basis for all results (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Nelson, Krapohl & Handler, 2008).

The four possible test results were No Significant Responses (NSR), Significant Responses (SR), Inconclusive (INC), and No Opinion (NO). An NSR result indicated the data were interpretable and signified a conclusion the examinee was likely truthful to all the relevant questions. An SR result also indicated interpretable data but that the conclusion was the examinee was probably deceptive to at least one relevant question. An INC outcome meant that the data were scorable but that the scores fell short of the thresholds for either an NSR or SR decision. An NO would be rendered if the data were not scorable (e.g., highly erratic tracings, suspected countermeasures, the session was terminated early, etc.). A minimum of three test charts were required. If after three charts the results would be INC, examiners recorded a fourth or fifth chart in an attempt to garner sufficient scores for an NSR or SR. No more than five scorable charts were permitted.

The instruments on which the cases were conducted were all produced by the Lafayette Instrument Company, either models LX5000 or LX6. The DV group submitted 244 of the cases with the remaining 171 cases coming from the HV examiners. Polygraph techniques included the mixed-issue screening Air Force MGQT with either two or three relevant questions, and the British One-issue Screening Test. Being field cases, ground truth was largely unavailable. Given the very restricted and potentially biased ground truth confirmation information, there was no attempt to compare ground truth against the decisions made by the two types of voice.

Table 1 is a cross tabulation of techniques for examiners using either their own voices or the digital voice. Tests of proportions (Bruning & Kintz, 1997) found those who used their own voices tested with the AFMGQT technique significantly more often than those who used the digital voice ($z = 2.24$, $p < 0.05$) and tested less often with the BOST ($z = 3.87$, $p < 0.05$). Differences in the proportion of sessions that were terminated did not reach significance ($z = 1.92$, ns)

Table 1. Cross tabulation of cases in which the examiner tested with her own voice or the digitized voice for the testing techniques of the AFMGQT and BOST, and when the session was terminated before testing was completed.

	AFMGQT	BOST	Session Terminated
Human Voice	142	29	3
Digital Voice	174	71	0

Procedure

As a routine business practice the test results, testing technique and number of test charts for all cases submitted for quality control review are recorded in an Excel spreadsheet, along with an identifier to indicate which type of voice the examiners were using during the testing phase of the examination. The present analyses used only those records. The cases were sorted and tallied for the present investigation.

As an archival study, there were no manipulation of variables or changes in testing procedures. Though of potential interest, neither examiner nor examinee opinions regarding their attitudes about the voices used in testing could be captured.

Results

Regarding a possible relationship between the testing voice and the test results, tests of proportion found no significant differences for any of the four test outcomes. See Table 2.

Table 2. Number and (proportions) of NSR, Inconclusive, No Opinion and SR results for polygraph cases in which either the examiner or the computer presented the test questions during testing. No significant differences were found between the type of voice for any of the examiner decisions.

	NSR	INC	NO	SR
Human Voice	77 (0.45)	22 (0.13)	12 (0.07)	60 (0.35)
Digital Voice	125 (0.51)	37 (0.15)	10 (0.04)	72 (0.30)

We also tried to determine whether the voice used during testing was related to the examiner's decision to record more data to reach a conclusion. Table 3 shows the frequency and (proportions) of cases that used 2, 3, 4 and 5 charts for the HV and DV methods. We compared each of the vertical columns using tests of proportions (Bruning & Kintz, 1997). Only one significant difference was found. It was the number of cases in which the session concluded before a minimum of three charts were recorded. The two-chart cases (early termination of the examination) occurred more often when the human voice was used than when the digital voice had been employed. See Table 3.

Table 3. Number and (proportions) of cases with 2, 3, 4 or 5 charts in which either the examiner or the computer presented the test questions during testing. * indicates a statistical difference at p < 0.05 between HV and DV.

	Number of Charts			
	2	3	4	5
Human Voice	3 (0.02)*	114 (0.67)	21 (0.12)	33 (0.19)
Digital Voice	0 (0.00)*	162 (0.66)	38 (0.16)	44 (0.18)

Discussion

Most polygraph schools teach examiners to present the test questions themselves rather than to rely on automation to manage that task. Most field examiners likewise prefer to use their own voices to present the test questions. Given the advantages discussed earli-

er in this article it remains unclear why this simple task has not been more widely given to automation.

Our evaluation of field data found the type of voice, human or automated, had no meaningful effect on the proportions of NSR, SR, INC or NO results. These null findings are unsurprising and may give reassurance to examiners who are considering implementing the automated voice in their own testing practices. Similarly, the number of test charts deemed necessary by the testing examiners was largely unaffected by the type of voice they used. The sole difference was found for the proportion of cases where the session was terminated after only two test charts, and for which the human voice accounted for more than did the digital voice. While of statistical significance, the difference consisted only of three cases in the HV condition and none in the DV condition. Also, there were eight separate comparisons made between the digital and human voice outcomes. Finding a significant result was more likely simply because so many comparisons were made. Further studies are encouraged to resolve whether the present findings are robust. In addition, it would be of interest and of practical value to solicit examiner and examinee attitudes about the voices used during testing.

Limitation

The source of the data in this study did not permit analyses as to whether the human or digital voice produced higher accuracy.

References

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.

Bruning, J.L., and Kintz, B.L. (1997). *Computational Handbook of Statistics* (4th ed.). New York, NY: Addison-Wesley (pp. 285 – 288).

Handler, M., Nelson, R., Goodson, W, and Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. P*olygraph 39*(4), 200 – 215.

Honts, C.R., and Amato, S. (2007). Automation of a screening polygraph test increases accuracy. *Psychology, Crime & Law, 13*(2), 187 – 199.

Kubis, J.F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations*. Project No. 5534, Task No. 553401, Prepared for Rome Air Development Center, Air Force Systems Command.

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.

Raskin, D.C., and Kircher, J.C. (1990, May 2). *Development of a Computerized Polygraph System and Physiological Measures for Detection of Deception and Countermeasures: A Pilot Study.* Preliminary Report. Contract 88-L655300-000. Scientific Assessment Technologies, Inc. Salt Lake City, UT.