EUROPEAN
POLYGRAPH
PUBLISHED SEMI-ANNUALLY
E-ISSN 2380-0550 ISSN 1898-5238
2026 VOLUME 20 NUMBER 1 (63)

https://doi.org/10.31749/2380-0550-EP2026-1-02

Same Question, Different Scores: Exploration of Two Possible Influences on the Variability between Spot Scores for the BOST

Donald J. Krapohl*
Donald Grubin**
Ian Dersley***

Abstract

The British One-issue Screening Test is a single-issue polygraph method intended for screening examinations. Although its two relevant questions are virtually the same we have found there to be a large average difference between the total scores for each. We undertook a 21-month study to identify possible contributors to this finding. In one condition examiners used two identical relevant questions. In the other condition examiners used slightly different verbiage

* Don Krapohl is a Past President of the American Polygraph Association and a regular contributor to this journal. Comments and suggestions should be sent to him at APAkrapohl@gmail.com
** Don Grubin is Emeritus Professor of Forensic Psychiatry at Newcastle University, England.
*** Ian Dersley is a PCSOT polygraph examiner in the Yorkshire and Humber and North East Regions of the UK National Probation Service. The authors express their gratitude to the examiners who provided the data used in this study: James Cook, Michael Reddish, Tracey Little, James Fraser, Harriet Gregan, Brad Hughes, Ian Dersley, Jo Scully, Bernard Morris and Caroline Perrell. This project would not have been possible without their assistance. The views expressed are solely those of the authors and do not necessarily represent those of their respective employers or affiliations. Conflict of Interest: The authors declare they have no conflicts of interest in this study.

for the two relevant questions, but the pretest instructions were modified. Using a large sample of control cases, we found no differences between the control samples compared with the use of two identical relevant questions, but there was a significant effect when there was a change in pretest instructions. Implications of this finding are discussed, including possible effects of priming and habituation.

Key words: British One-issue Screening Test, Empirical Scoring System, score variability

From a psychophysiological viewpoint, polygraph examinations are a test of salience, which is the trigger that evokes physiological responding. Salience itself may be primed by emotional and cognitive factors (see Khan et al., 2009). Relative response intensities observed in polygraph recordings provide an index of the personal significance of the test questions – the more significant, the larger the reaction on average. Polygraph testing takes advantage of this tendency to make inferences regarding an examinee's veracity: Stronger reactions to relevant questions have an established relationship with deception, while stronger reactions to comparison questions are associated with truthfulness (Nelson, 2015).

If polygraphy is a test of salience, a related assumption is that relevant questions which share the same content should evoke physiological responses of similar intensity. In a review of 180 manual scores assigned to relevant questions on British One-issue Screening Tests (BOST) carried out in 2022, a test where the two relevant questions cover identical behaviors and time periods, we found an average difference between the total scores of each relevant question to be 4.4 points (sd = 3.17), with a range of 0 to 17 points (unpublished). The correlation between total scores of each of the two relevant questions was significant (Pearson's $r = 0.35$) but small (using Cohen's 1988 interpretation). This modest correlation and the relatively wide range of score differences were not anticipated given that the two relevant questions were virtually identical.

We had two working hypotheses to explain these findings. The first concerned the characteristics of the test questions themselves. We proposed that though the relevant questions are merely a slight rewording of one central question, examinees may attach different levels of salience idiosyncratically, depending on the verbiage of the individual questions. A second hypothesis is that something in the pretest instructions, yet to be identified, differentially affected the salience examinees assigned to the two relevant questions.

To test the first hypothesis, we devised a field study in which a group of examiners used identical wordings for both relevant questions. To test the second, we assigned a group of examiners to use the pretest instructions of one of the examiners whose two scores historically differed less than average (Appendix A).

We also created a control group of examiners who knew they were participating in the study but were advised to continue their current pretest and testing practices. Mindful of the Hawthorne Effect, where behaviors can change merely because they are being observed, we would compare those data with our 2022 findings, and if consistent with them we would combine them to create a larger control group.

To maximize sample sizes we tracked the scores of all BOST cases for the three groups over 21 months.

Method

Examiners and Instructions

We recruited 10 volunteers from among the 46 polygraph examiners in a large offender management program in the UK. All had attended the same APA-accredited polygraph education program, had at least three years of experience, conducted more than 100 field examinations, and participated in continuing education offerings three times per year.

Three groups were created using a random number generator. Group 1 consisted of three examiners whose instructions were to use identical wording for both relevant questions when conducting BOST examinations. Group 2 had four examiners who were provided with a script with which to introduce the two relevant questions in their BOST examinations (Appendix A). The remaining three examiners comprised the control group and were instructed to conduct their BOST examinations without any changes.

Cases

All examinations were conducted on Lafayette computer polygraphs LX5000 or LX6. The recorded data were electrodermal, cardiovascular, vasomotor, a motion sensor and two breathing channels. The study data consisted of all BOST cases conducted by the 10 volunteers between April 1, 2023, and December 31, 2024.

Because the BOST examinations were field cases, ground truth regarding the examinee's veracity was unavailable.

Group 1 conducted 50 BOST cases. Two cases were excluded, one because the examinee failed to follow the examiner's instructions and the second case due to concerns regarding the examinee's mental health. This left 48 examinations for analysis of which 35 resulted in a decision of No Significant Responses (NSR) and 10 of Significant Responses (SR) results, with the remaining three deemed Inconclusive (INC).

Group 2 conducted 34 BOST cases, all of which were included in the study. There were 25 NSR examinations, six SR, and three INC.

The control group conducted 46 BOST cases. One case was removed due to poor quality data attributed to the examinee's health. Of the remaining 45 cases, 28 resulted in NSR, 16 in SR and one as INC.

We compared the control group data to the 2022 data that prompted this study. None of the proportions, average differences or correlations were found to be significantly different between the study data and the previous data (Table I). This supported a conclusion that the data from the study control group was similar to the baseline data of 2022. Consequently, the control group data and the 2022 data were combined to create a larger control group consisting of 225 cases.

Table I. Summary statistics of BOST results and scores comparing the control group with the entire data set from 2022

	Study Data	2022 Data
Sample size	45	180
Proportion of NSR	0.62	0.66
Proportion of SR	0.36	0.28
Proportion of INC	0.02	0.07
Avg difference (s.d.) between R1 and R2 scores	4.31 (3.73)	4.42 (3.17)
Correlation between R1 and R2 scores	0.36	0.35

British One-issue Screening Test

The BOST has been previously described in detail elsewhere (Krapohl et al., 2020). Briefly, the BOST has the same question types and sequence as Variation 1 of the

Air Force Modified General Question Test (AFMGQT) with two relevant questions (Krapohl & Shaw, 2015). However, it differs from the AFMGQT in two meaningful ways. One is that the two relevant questions in the BOST must encompass identical behaviors and time periods such that an examinee must be either truthful or deceptive to both questions, in contrast to the AFMGQT which has no similar constraint. Second, the decision rules for the BOST are based on the sum of all scores whereas decision rules for the AFMGQT consider only the sum of scores for the individual questions (see the next section for details regarding scoring and decision rules). As practiced in this program, all relevant and comparison questions are systematically rotated in the question sequence during the testing phase.

Scoring System

All cases were manually scored by the examiners using the Empirical Scoring System (Nelson et al, 2011). Physiological responses for each relevant question were scored against the stronger response to the closest comparison question presented before the relevant question or the one after it. At the end of testing the scores were summed for each of the relevant questions and for the entire test.

BOST decisions of NSR require that the total of all scores for the entire examination sum to +2 or more. If the total score is -4 or lower the decision is SR. If the results of the examination would be inconclusive, two-stage rules are imposed (Senter, 2003) which require a decision of SR if the sum of scores of either relevant question is -6 or lower. All other results remain Inconclusive.

Procedure

The scores and decisions assigned by the participating examiners were recorded in an Excel spreadsheet. Means, standard deviations and correlations were calculated using applications within the Excel program. Online calculators were used for tests for differences between means and differences in correlations (Pearson's r)*,**. For significance testing of differences between proportions the first author developed a computational sheet in Excel using a statistical formula found in Bruning and Kintz (1997). Alpha was set at .05 for all comparisons. Because this was an exploratory study we made no Bonferroni corrections so we would be able to identify subset of potentially significant effects that could be studied in subsequent research.

---

* https://www.medcalc.org/calc/comparison_of_means.php

** https://www.danielsoper.com/statcalc/calculator.aspx

Results

Group 1

Table II shows descriptive statistics for the group of cases in which two identical relevant questions were used in the BOST and the corresponding statistics for the now-larger control group. There were no significant differences found between the two groups.

Table II. Summary statistics for BOST results and scores comparing Group 2 with the control group. No comparisons were statistically significant at $p = .05$

	Group 1	Control Group
Sample size	48	225
Proportion of NSR	0.73	0.65
Proportion of SR	0.21	0.29
Proportion of INC	0.06	0.06
Avg difference (s.d.) in RQ scores	3.96 (2.92)	4.41 (3.28)
Correlation between R1 and R2 scores	0.53	0.35

Group 2

Table III shows descriptive statistics for the group of cases in which there were changes to the pretest introduction of the relevant questions in the BOST compared with the corresponding control group statistics. There were no significant differences found between Group 2 and the control group except for the correlation between the scores of the two relevant questions where there was a significant difference between Group 2 and the control group, with a stronger correlation in the former ($z = 2.52, p < .05$).

Table III. Summary statistics of BOST results and scores comparing Group 2 with the control group. The * indicates a significant difference between Group 2 and the Control Group

	Group 2	Control Group
Sample size	34	225
Proportion of NSR	0.74	0.65
Proportion of SR	0.18	0.29
Proportion of INC	0.09	0.06
Avg difference (s.d.) in RQ scores	3.29 (3.18)	4.41 (3.28)
Correlation between R1 and R2 scores	0.69*	0.35*

Discussion

We found no differences between the control group and the condition in which two identical relevant questions were used in the BOST, even with the increased risk of a Type I (false positive) error resulting from multiple comparisons without Bonferroni corrections. The present data do not support a hypothesis that differences in BOST scores between the two relevant questions are associated with different wordings of the questions. Our findings suggest that it may make no difference whether examiners choose to use identical or slightly different verbiage for the two questions in the BOST, and that rules endorsing one approach over the other may be unnecessary but would benefit from further investigation.

Compared to the control group, the scripted instruction used by Group 2 did not result in any differences in the proportion of test outcomes, nor for the average difference in scores between the first and second relevant question. There was, however, a significant difference in the correlation coefficients for the scores of the first and second relevant questions, with Group 2 having a significantly higher correlation between the scores of the two relevant questions than the control group. This is a mixed finding that defies easy interpretation.

It appears that the verbiage of the test questions may have limited influence on the differences in test scores inasmuch as neither group showed a reduction in the differences in scores between the two relevant questions compared to the control group. But the finding of a stronger correlation between scores for the two relevant questions for Group 2 compared with the control group has two possible explanations. One is simply a Type 1 error, that is, it is a false positive finding. The second is that the changes in the pretest instructions used by Group 2 increased the coupling between the scores of the two relevant questions such that the two sets of scores tended to move in the same direction more consistently than did those of the control group even though there was still a difference in the absolute response. Said another way, though the average difference in the scores between the first and second relevant question seem to be unaffected by how they are phrased, the unique pretest instructions are associated with a tendency for the two scores to follow each other more closely than do the two scores from the control condition.

One of the concerns expressed by some examiners is that occasionally the score of one relevant question in the BOST is of the opposite sign to the score of the other question, despite the fact that they both encompass identical behaviors and time periods. This means that one question could be trending toward deception while

the other is toward truthfulness, the decision then being based on which of the two scores pulls the grand total score over a decision threshold. In a *post hoc* analysis we looked at the frequency for such opposing scores. This occurred in 26%, 23% and 15% of cases for the control group, Group 1 and Group 2, respectively. Tests of proportions (Bruning & Kintz, 1997) found no significant differences in proportions of opposing scores among the groups. Though it may be disconcerting to examiners, the opposite signs may not be as large a problem as it seems. Though occurring in a minority of tests this phenomenon is not uncommon. It may be the result of a habituation effect in cases where there is a lesser response the second time the question is asked, or to a priming effect when the response is greater in the second question, with variation between examinees. This possibility will be explored in a planned replication. Regardless, the use of the grand total score in BOST decision-making would be expected to provide a more stable estimate of the examinee's veracity than the scores of individual questions due to the differences in the number of samples available between the two approaches. The differences in total scores for the two relevant questions may simply be the manifestation of the curse attendant to smaller sample sizes (Lindstromberg, 2023).

It remains the case that the reason for different response intensities to virtually identical questions in the BOST remains a puzzle. We believe that further research exploring this, especially the impact of habituation and priming effects, would be of benefit, not only in explaining our findings with the BOST, but also contributing to a better understanding of the psychology underlying polygraph testing.

Limitations

Ground truth was not available for any but a small minority of cases. Therefore, it was not possible to test for differences in validity among the different conditions in this study.

The manual ESS method used in the offender management program from which the cases were drawn may be different from that used elsewhere. The program has implemented empirically based ESS scoring rules regarding onset latencies (Krapohl et al., 2021) and minimum response amplitudes (Krapohl et al., 2019) that are not universally practiced.

We acknowledge the inflated alpha that is a consequence of our multiple comparisons without Bonferroni corrections. Our chosen statistical approach gives greater confidence to the many null findings than to the finding of a singular significant

effect. Our approach served its purpose in reducing the scope of our planned follow-on investigation, but it did not provide concrete evidence that the single apparent effect is genuine.

References

Bruning, J.L., & Kintz, B.L. (1997). *Computational Handbook of Statistics* (4th Ed.). Addison Wesley Longman: Reading, MA.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Khan, J., Nelson, R., and Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph, 38*(3), 184–197.

Krapohl, D.J., & Dutton, D.W. (2022). A field assessment of manually scoring electrodermal data in self-centering and non-centering modes. *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice, 51*(1), 20–30.

Krapohl, D.J., Dutton, D.W., and Nix, K.A. (2019). A brief discussion of the lower latency limit of the electrodermal response in polygraph testing. *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice, 48*(2), 98–104.

Krapohl, D.J., Grubin, D., Benson, T., and Morris, B. (2020). Modification of the AF-MGQT to accommodate single-issue screening: The British One-issue Screening Test. *Polygraph & Forensic Credibility Assessment: Journal of Science and Field Practice, 49*(2), 176–183.

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

Lindstromberg, S. (2023). The winner's curse and related perils of low statistical power – spelled out and illustrated. *Research Methods in Applied Linguistics*, 2(3), https://doi.org/10.1016/j.rmal.2023.100059

Nelson, R. (2015). Scientific basis for polygraph testing. *Polygraph, 44*(1), 28–61.

Nelson, R., Blalock, B., and Handler, M. (2011). Criterion validity of the Empirical Scoring System and the Objective Scoring System, version 3 with the USAF Modified General Question Technique. *Polygraph, 40*(3), 172–179.

Senter, S.M. (2003). Modified question test decision rule exploration. *Polygraph, 32*(4), 251–263.

Appendix A. Example pretest interview script for BOST relevant questions

How to scope the RQs in BOST tests

Example questions:

Since we last met, have you had any contact with an under 18-year-old?

Since I last saw you, have you had any contact with someone aged under 18?

'Okay Fred, I'm going to look first at the questions that relate specifically to your licence conditions*. The first one I want to look at is, 'Regarding your licence conditions, do you intend to answer each question truthfully?'

Any discussion that then takes place about the Sacrifice Relevant Question.

Now as far as the other questions about your licence conditions are concerned, your probation officer has only asked me to look at one area with you.

That means I'll have to do something slightly different to the way in which I'd normally do a test because polygraph tests weren't originally designed to look at just one area. The way we get round that is by asking one version of a question and then we ask another version of the same question using slightly different wording. But it means the same thing. So when I'm collecting the charts, it will sound as though I'm asking you about the same thing, twice. It's not a trick or anything that is trying to catch you out. If it was, I wouldn't be telling you about it (hearty laugh!).

So the first version of the question will be, 'Since we last met, have you had any contact with an under 18 year old?'

By 'contact' I mean any face-to-face interaction, including via any type of video app; any contact via a phone – so texts and telephone calls, contact via any type of internet enabled device – so smart phones, laptops, tower computers, xBoxes, Playstations, palms, tablets, smart watches or any other type of internet enabled device you can think of.

There we are talking about contact via websites including dating and contact sites, social media, online forums, chatrooms or message boards – such as WhatsApp, emails or any other type of platform that you have to access the internet to use.

---

* The expression "licence conditions" in the UK would mean the same as terms of probation or parole in the US.

It also includes contact in writing – letters, notes, birthday card, Christmas card or anything else that is written or typed and sent or given to someone aged under 18.

Lastly, it includes any contact that takes to form of passing or receiving verbal messages through third parties.

So – if I ask you that question, 'Since we last met, have you had any contact with an under 18-year-old?, what will your answer be?'

Fred: – No

Me: – 'Okay, that's great. So, the second version of the question will be, 'Since I last saw you, have you had any contact with someone aged under 18?' Now, obviously, that means exactly the same thing as the first version of the question. So, would I be right in assuming that your answer to it will be the same as it was to the other version of the question?

Fred: – Yes.

Me: – 'Great – so your answer will be .....?

Fred: – No