# Short Report:

**Bootstrap Analysis of Serial Testing Effects Using Laboratory Data**

# Raymond Nelson

**ABSTRACT**

Criterion accuracy was analyzed for the *serial-single-issue-screening-test* (SSIST) using laboratory data from two previous publications in which the test directed-lie-screening-test (DLST) format was adapted to the single-issue testing context. The serial testing context involves the aggregation of several single-issue exams – each with multiple presentations of a set of non-independent RQs, together with other procedural questions – into a single examination. SSIST outcomes are negative

– requiring no further interviewing, testing, or investigation – when the results are negative for all series. SSIST results are positive – which may require additional interviewing, testing, or investigation, when the results of any series are positive. Previous results using Monte Carlo methods suggested that serial testing strategies do not prevent well-known statistical *multiplicity* effects – also known as the *inflated-alpha* effect (Nelson, 2023). These are also referred to as a *cumulative error* effect, and more generally as the *multiple comparison* effect or the *problem of multiple comparisons.* Analytic results from this project further illustrate the cumulative error effects that can result from the use of serial testing strategies and highlights the need for additional research into the computation and management of cumulative error effects for both multiple-issue and SSIST methodologies.

The serial-single-issue-screening-test (SSIST) is a multiple-issue screening approach that attempts to leverage the criterion accuracy effects of a single-issue test format in the multiple-issue testing context1. O’Burke et al., (2022) proposed the use of a single-issue single-sequence (i.e., several iterations of the test questions in one recorded sequence) screening test and hypothesized that a series of single-issue screening tests may be advantageous compared to multiple-issue polygraph screening test formats.

O’Burke (2022) described the use of the SSIST approach with each target issue within a multiple-series exam. They described each series (identified by an alphabetic character) as addressing a different test target issue: series-A illegal drug use, series-B serious crimes, series-C domestic violence, and series-D sex crimes. Examinees who pass all series are classified as truthful or passing the examination (i.e., test results are reported as *no-significant responses not indicative of deception*). Examinees are classified as deceptive (i.e., test results are reported as *significant responses indicative of deception*) if they produce positive result for any series.

Analysis of the SSIST approach using Monte Carlo methods (Nelson, 2023) suggested the hypothesized advantages may be elusive, because serial testing strategies do not prevent well-known statistical *multiplicity* effects – also known as the *inflated-alpha* effect, or *cumulative error* effect, and more generally as the *multiple comparison* effect or the *problem of multiple comparisons.*

Multiplicity effects are the reasons that researchers and statisticians are attentive to the potential problems that surround the calculation of multiple statistical outcomes – including multiplicity effects stemming from a primary analysis and those surrounding *post- hoc* analysis (referred to by polygraph field practitioners as *break-out* testing) in general. Knowledge and experience associated with multiplicity effects is the basis for routine use of omnibus analytic method (i.e., the family of ANOVAs in scientific research, and multiple- issue test formats in polygraph screening contexts) when possible, and the use of statistical corrections (e.g., Bonferroni and Sidak) when practical [See Nelson (2015) for an introduction and discussion.]

Previous analysis using Monte Carlo methods (Nelson, 2023) provided some insight into the expected effects of the SSIST approach compared to the traditional omnibus (i.e., multiple-issue) testing approach. Results from the Monte Carlo analysis revealed an increase in test sensitivity to deception and decrease in false-negative errors, along with a disproportionate decrease in test specificity with corresponding increases in false-positive errors and inconclusive results. The aggregated effects showed a decrease in overall classification accuracy compared to the multiple-issue testing effects described by APA (2011). This project is intended to further explore the expected effects, and to supplement the growing basis of scientific evidence and knowledge, associated with the SSIST.

# Method

Data were obtained from two laboratory studies (Prado, et al., 2015a; 2015b) on single-issue polygraphs consisting of two non-independent relevant questions (RQs) presented four times in a single sequence. The diagnostic test formats described in the Prado et al. studies are an adaptation of the multiple-issue *directed-lie-screening- test* (DLST). Although the name was not included in the two publications, this single- issue test format is sometimes referred to by polygraph field practitioners as the *directed- lie-diagnostic-test* (DLDT)2. Scores from the two laboratory studies were aggregated and subject to bootstrap resampling to model the effects of a series of single-issue screening polygraphs. The mean score for criterion guilty cases was -3.67 (sd=5.76) and the mean score for criterion innocent cases was 5.26 (sd=5.48). The bootstrap model allowed for the criterion state of each series (each target issue) to be known with certainty – because it was established in the controlled laboratory experiment – and independent criterion states were used for each series in each case within the bootstrap model3.

1 Diagnostic testing differs from screening in the presence of a known problem (i.e., allegation or incident) for which a person is suspected of involvement. Screening polygraph tests differ from diagnostic exams in one other important way

– screening polygraphs are often conducted as multiple-issue exams and are interpreted with an assumption of non- independence (APA, 2011; 2023).

# Analysis

Bootstrap resampling methods were used to calculate statistical confidence intervals for test sensitivity (true-positive or TP) and specificity (true-negative or TN) rates in addition to false-negative (FN) and false-positive (FP) error rates. These statistics were then used to calculate additional criterion accuracy metrics, including the inconclusive rates for criterion-innocent (I-INC) and criterion-guilty (G-INC) cases, the proportion of correct (non- inconclusive) classifications for criterion-guilty (GPC) and criterion-innocent cases (IPC) , the unweighted accuracy (ACCY) and unweighted inconclusive (INC) rates. Positive-predictive- value (PPV) and negative-predicative-value (NPV) were also computed. A double bootstrap procedure, consisting of 1 million iterations of the sampling data, was used to calculate the standard errors, and confidence intervals were calculated as the .025 and .975 quantiles of the bootstrap space.

# Results

Table 1 shows the point estimates and 95% confidence intervals for accuracy effects using the aggregated data from the Prado et al. studies, along with SSIST models that include two, three and four independent target issues. In the SSIST approach each separate target issue is presented as a separate test series. Each series consists of several presentations/ repetitions of the RQs and other test questions. For single-issue exams the RQs are interpreted with an assumption that they are non-independent. However, a multi-series exam may consist of several different sets of test questions. A complete SSIST consists of all series (targets). Examinees are said to produce negative test results if the results are negative (i.e., statistically significant for truth-telling) for all question series, and are said to produce positive results if the results are positive (i.e., statistically significant for deception) to any question series. The overall test result is correct if the results for all series are correct, and the test result is incorrect if the result of any series is incorrect. The overall test result is inconclusive if none of the series is positive and one or more series is not statistically significant for truth-telling. In this way, interpretation of SSIST outcomes is conceptually similar to multiple-issue polygraph formats. Also shown in Table 1 are the point estimates and confidence intervals for multiple-issue polygraph techniques (APA, 2011), and for the aggregated results from the published studies on the DLST (Prado et al., 2015a; 2015b).

2 Diagnostic tests are those tests conducted in response to a known problem (APA, 2023), and make use of single-issue test formats for which multiplicity problems can be avoided or easily managed. Screening tests are those tests conducted in the absence of any known problem (e.g., known incident or known allegation). Although the practical goals may differ, validation of both diagnostic and screening tests is achieved through analysis of test sensitivity, specificity, and error rates. When these are known with reasonable confidence it is acceptable to generalize expected effect sizes from one use context to another.

3Multiple criterion states for field polygraph are interpreted with an assumption of independence but are known to be incompletely independent. There are a variety of factors that potentially limit the independence among multiple-issue polygraph targets, including the limited attentional resources of the examinee and the fact that some behaviors be covary (e.g., serious crimes, sex assault).

**Table 2. Accuracy summary for serial single-issue exams with independent (mixed) criterion**

**states, including Mean (SE) [95% CI].**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | DLDT(Aggregated data from Prado et al, 2015a, 2015b) | SSIST2 separate series (2 Target Issues\*) | SSIST3 separate series(3 Target Issues\*) | SSIST4 separate series(4 Target Issues\*) | Multiple Issue PDD Techniques (APA, 2011) |
| Sensitivity (TP) | .717 (.065) | .751 (.061) | .762 (.060) | .767 (.059) | .771 (.072) |
|  | [.592 to .836] | [.630 to .859] | [.642 to .868] | [.648 to .871] | [.630 to .911] |
| Specificity (TN) | .779 (.059) | .605 (.070) | .471 (.071) | .366 (.069) | .719 (.047) |
|  | [.659 to .889] | [.468 to .736] | [.337 to .605] | [.235 to .497] | [.626 to .811] |
| False negative (FN) | .154 (.052) | .139 (.049) | .132 (.048) | .131 (.048) | .113 (.058) |
|  | [.061 to .264] | [.053 to .238] | [.048 to .228] | [.047 to .226] | [.001 to .226] |
| False positive (FP) | .085 (.039) | .163 (.053) | .235 (.060) | .300 (.065) | .144 (.039) |
|  | [.019 to .167] | [.071 to .267] | [.128 to .352] | [.184 to .428] | [.066 to .221] |
| Guilty inconclusive (G-INC) | .129 (.047) | .111 (.044) | .106 (.043) | .102 (.042) | .112 (.051) |
|  | [.041 to .228] | [.034 to .200] | [.031 to .194] | [.029 to .189] | [.013 to .212] |
| Innocent- inconclusive (I-INC) | .136 (.048) | .232 (.061) | .294 (.065) | .334 (.068) | .136 (.031) |
|  | [.048 to .235] | [.124 to .351] | [.175 to .421] | [.209 to .446] | [.076 to .196] |
| Unweighted inconclusive (INC) | .132 (.034) | .171 (.038) | .200 (.039) | .218 (.040) | .125 (.029) |
|  | [.070 to .200] | [.104 to .246] | [.128 to .277] | [.143 to .297] | [.068 to .183] |
| Guilty percent correct (GPC) | .823 (.058) | .844 (.055) | .852 (.053) | .855 (.053) | .873 (.066) |
|  | [..700 to .929] | [..734 to .940] | [.747 to .946] | [..749 to .948] | [.744 to .999] |
| Innocent percent correct (IPC) | .902 (.046) | .788 (.067 | .667 (.079 | .549 (.087) | .831 (.043) |
|  | [.805 to .978] | [.654 to .905] | [.512 to .812] | [.381 to .549] | [.746 to .915] |
| Unweighted accuracy (ACCY) | .862 (.037) | .816 (.044) | .760 (.048) | .702 (.051) | .850 (.039) |
|  | [.786 to .930] | [.729 to .893] | [.666 to .848] | [.603 to .796] | [.773 to .926] |
| Positive predictive value (PPV) | .894 (.049) | .822 (.057) | .765 (.060) | .719 (.062) | .828 (.059) |
|  | [.789 to .976] | [.706 to .921] | [.646 to .873] | [.599 to .833] | [.712 to .943] |
| Negative predictive value (NPV) | .836 (.054) | .813 (.065) | .780 (.077) | .737 (.090) | .878 (.049) |
|  | [.718 to .933] | [.683 to .927] | [.628 to .917] | [.558 to .899] | [.782 to .973] |

\* Prior probabilities were calculated using the Sidak equation 1-(1-a)^(1/n) to hold the aggregated prior base rate constant at .5 (1 to 1) to facilitate an intuitive contrast between the 2RQ, 3RQ, and 4RQ conditions.

# Discussion

Results from this project, using data from a controlled laboratory study, further illustrate the potential for multiplicity effects with the SSIST, leading to overall criterion accuracy rates that differ markedly from the accuracy of both multiple-issue polygraph exams and single-series single-issue polygraphs. These results, together with previously reported results from Nelson (2023), suggest that serial testing strategies can introduce cumulative error problems, and do not offer a convenient and simple solution to the complex problem of multiple-issue polygraph screening. Previous analyses (Barland, Honts & Barger, 1989; Nelson, 2023) suggest that cumulative error effects are not avoided by this approach but are instead merely relocated. Data in this analysis show that overall accuracy may decrease as a result of this approach and suggest the need for further exploration and study of before embarking on the wide-scale replacement of multiple-issue polygraph test formats for which statistical corrections can be conveniently applied to the cutscores and analysis methods.

Two main factors influence test accuracy: 1) whether the test target issues are assumed to be independent or non-independent, and

2) the number of RQs. (Nelson & Handler 2017). Importantly, use of more RQs can increase the precision or accuracy of single- issue exams but can decrease the precision or accuracy of multiple-issue exams. The present study addresses the ambiguity of the SSIST – whether the SSIST can maintain the accuracy of single-issue polygraphs in the multiple-issue testing context? Evidence at this time has not been published to support the idea that this is possible. Instead, available evidence suggests that serial single-issue polygraphs share well- known statistical phenomena with multiple- issue exams. Serial single-issue polygraphs may be subject to cumulative error effects similar to those of multiple-issue exams with corresponding decreases in test accuracy compared to the use of a single-series single- issue exam. There are some contexts in which the increase in test sensitivity and reduction of false-negative errors may be a desirable effect, even when considering the reduction of test specificity and decrease in overall accuracy.

Single-issue screening tests are not a new idea within the polygraph profession (Barland, Honts, & Barger, 1989), and other testing contexts. Other possibilities for a single-issue screening polygraphs have also been proposed (Barland, Honts, & Barger, 1989; Krapohl et al., 20204). Interest among field practitioners and program managers warrants further research on the SSIST, and other single-issue polygraph screening formats, using controlled laboratory studies for which the criterion can be known for multiple test targets. Future research should address the issues surrounding the selection of appropriate prior probabilities, the impact of different priors on posterior probabilities, and the potential use of statistical corrections to manage cumulative error effects. More information is needed to understand the different advantages and disadvantages of the SSIST approach compared to more common omnibus testing and analysis methods. Although the goal of the SSIST may be to leverage the accuracy effect sizes of a single-issue test format in the multiple-issue screening context, data at this time indicate that the accuracy effect sizes for multiple-series exams are not equivalent to the effect sizes for single-series single-issue exams.

4 The important difference between the single-issue screening test described by Krapohl et al., 2020 and the SSIST is that Krapohl et al. propose the use of a single-series single-issue screening approach, which avoid the potential for cumulative error effects.

# References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated

polygraph techniques. *Polygraph, 2011, 40(4):* 195-305. [Electronic version] Retrieved

January 15, 2023, from [http://www.polygraph.org.](http://www.polygraph.org/)

American Polygraph Association. (Aug. 25, 2023). *APA Standards of Practice.* Retrieved from [https://www.polygraph.or](http://www.polygraph.org/)g on Nov. 7, 2023.

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). *Studies of the Accuracy of Security Screening Polygraph Examinations.* Department of Defense Polygraph Institute.

Handler, M., Nelson, R. & Blalock, B. (2008). A focused polygraph technique for PCSOT and law

enforcement screening programs. *Polygraph, 37(2):* 100-111.

Krapohl, D. J. Grubin, D., Benson, T., & Morris B. (2020). ModiÀcation of the AFMGQT to accommodate single-issue screening. the British one-issue screening test. *Polygraph & Forensic Credibility Assessment, 49(2):* 176-183.

Nelson, R. (2015). Bonferroni and Sidak corrections for multiplicity effects with subtotal scores of

comparison question polygraph tests. *Polygraph, 44(2):* 162-167.

Nelson, R. (2023). Multiplicity Effects in the Serial Single-issue Testing Situation: When is a Single-issue Test Not a Single-issue Test? *Polygraph & Forensic Credibility Assessment, 52(2):* 129-147.

Nelson, R., & Handler, M. (2017). Practical polygraph: how to select a polygraph test format.

O’Burke, J. P., Westerman, D., & Goodson, W. (2022). Using single-issues in screening examinations. *APA Magazine, 55(6):* 51-59.

Prado. R., Grajales, C., & Nelson, R. (2015). Laboratory study of directed lie polygraphs with Spanish speaking examinees. *Polygraph, 44(1):* 79-90

Prado. R., Grajales, C., & Nelson, R. (2015). Laboratory study of a diagnostic polygraph technique in a single sequence: a replication study. *Polygraph, 44(1):* 1-12.