Multiplicity Effects in the Serial Single-issue Testing Situation: When is a Single-issue Test Not a Single-issue Test?

Raymond Nelson

Abstract

The serial-single-issue (SSI) approach to multiple-issue polygraph screening was investigated with the goal of understanding potential multiplicity effects that may play a role in effect sizes and expected advantages associated with this method. Using sensitivity, specificity and error rates from previously published polygraph studies, Monte Carlo methods were used to study the SSI situation. Examples are provided for the multiple-issue polygraph screening context and repeated testing in the single-issue diagnostic polygraph context. To provide additional insight around potential advantages of serial testing strategies, repeated testing strategies are also modeled for the COVID-19 context. Results indicate that serial testing can potentially decrease overall test accuracy with a disproportional increase in false positive errors and increased inconclusive results among innocent persons. Serial testing may provide smaller improvements in test sensitivity and false-negative errors. Expected accuracy and error rates are shown as a previse for readers who may wish to better understand the potential advantages and disadvantages associated with the use of a series of single-issue screening tests. Discussion of statistical multiplicity (cumulative error effects) and the use of omnibus analytic methods is provided. The greatest observed effect from the SSI approach was a large reduction of test specificity. Although some advantages may exist when testing one question or hypothesis at time, available evidence does not support the complete abandonment of omnibus analysis methods or multiple-issue polygraph screening techniques at this time.

Introduction

In a recent publication, O'Burke (2022) described the use of four screening target issues, referred to as relevant questions in polygraph field practice, including illegal drug use, involvement in serious crimes, domestic violence, and sex crimes, and has suggested the use a series of single-issue exams instead of the common practice of using multiple-issue screening tests. Previously published research (Barland *et al.*, 1989) did not show an advantage to the use of a series of single-issue exams when compared with the effects of a single multiple-issue exam. However, O'Burke has hypothesized that a serial-single-issue (SSI) screening approach, using the directed-lie-screening-test format (Handler *et al.*, 2008; Research Division Staff, 1995a; 1995b) as a single-issue format, as previously described by Prado *et al.* (2015a; 2015b)[1], may provide more accurate diagnosis of an examinee's problematic behavior, and may provide a more optimal use of available testing resources.

No analysis was provided by O'Burke (2022) to support the SSI hypothesis, and the only evidence discussed was an anecdotal description of unconfirmed cases[2] for which the single-issue test totals were observed to be greater than the subtotal scores that have been reported for multiple-issue screening tests[3], and for which the agency and examiners were reported anecdotally as satisfied. Means and standard deviations were reported for the observed truthful and deceptive outcomes. However, the included graphics showed that the data were severely non-normally distributed – such that any attempt to make practical or analytic use of these descriptive statistics using traditional parametric methods would lead to conclusions that are unstable and unreproducible.

In addition to the absence of supporting data and analysis, no information was provided for SSI effect sizes, and no discussion provided around the central fact that the SSI approach will be subject to the same known statistical phenomena as all other scientific and statistical activities involving multiple classification and inference activities. Finally, although economic effects may exist in terms of time, funding, physical space, personnel resources, or increased accuracy of deceptive classifications, O'Burke provided no econometric data or analysis to support the hypothesis of these effects.

A useful way to gain perspective and objectivity for a scientific hypothesis or idea is to find other examples that can be used to change the applied context. Valid constructs are expected to exhibit domain consistency, meaning that we can expect to observe similar phenomena and similar effects based on the concept regardless of the context.[4] Valid constructs are seldom valid in a single application paradigm. What other examples can be found for the use of a series of single-issue tests? In other words, if the SSI hypothesis is valid and useful, can we identify other examples of the use of serial testing strategies?

The recent global pandemic provides a potential example for the construct of domain consistency. Among the myriad of rules and guidelines that were offered during the COVID-19 pandemic was a suggestion that multiple consecutive negative test results might be used as a basis for certain decisions, such as ending a period of quarantine or isolation.[5]

---

[1]Prado et al. reported the use of a DLST format as a single-issue exam with three repetitions of 2 RQs in a single recording. They found accuracy rates similar to other single-issue polygraph techniques with 2 RQ, and no advantage to the use of a single chart recording.

[2]Presumptive guilt and innocence are not warranted for the subjects in this convenience sample because the evidence for such a presumption is not independent from the polygraph test result. Published knowledge on error rates for validated polygraph technique indicates that it should be reasonably presumed that the observed results include some proportion of testing errors.

[3]This finding is not surprising when considering that manual scores for polygraph examinations are aggregated via summation. Given the theory of the polygraph test, it is axiomatic that grand total scores are expected to be greater than subtotal scores.

[4]An example of this can be seen in personality and intelligence testing in the field of psychology, although deep and interesting discussion are ongoing as to the actual nature of these constructs, a construct - an abstract idea that is formulated to be thought of as somewhat tangible- can be accepted as valid when different contextual applications lead to similar observations, effects, and results. Although some differences are inevitable, intelligence and personality scores tend to correlate strongly from different methods of measurement and analysis.

[5]This recommendation was emphasized more strongly during the time period prior to the introduction of vaccines that substantially reduced the likelihood of severe illness, hospitalization, and death. It has been less talked about as the public has habituated to a more endemic form of viral illness.

A requirement for multiple COVID-19 test results – multiple single-issue testing – was an attempt to make deliberate use of the phenomena of *statistical multiplicity*, also referred to more simply as *multiplicity*, and the *problem of multiple comparisons*. When referring to positive test results the phenomena has been termed *inflated alpha* or the *cumulative error* effect because it describes the inflation of false positive error rates beyond a planned tolerance. When describing negative test results the phenomena of statistical multiplicity results in the opposite – a reduced or *deflated alpha* threshold. In the COVID-19 testing context, a requirement for multiple negative test results is expected to decrease the occurrence of false-negative errors.

Multiple-issue screening polygraph exams are a contextual allegory to the use of multiple comparison methods in scientific research. Decision rules for these exams can be thought of as allegorical to *omnibus* statistical analysis methods in which multiple hypothesis are tested in a single experiment and analysis. Results from scientific activities involving omnibus methods are often a useful and preferred solution in scientific research – even though they may require post-hoc analysis to better understand the resulting information when significant differences are observed.

A known limitation of omnibus methods, such as ANOVA´s, is that, although they may identify the presence of significant differences within an array of experimental conditions, they do not pinpoint the exact areas where differences occur. Multiple-issue polygraph examinations have been found to exhibit similar limitations (Barland *et al.*, 1989; Podlesney & Truslow, 1993; Raskin *et al.*, 1988). In field polygraph practice, *post-hoc* analysis is often referred to as *breakout testing or successive hurdles* testing, in which single-issue examinations may be conducted following a positive result from an omnibus or multiple issue-test.

Researchers, statisticians, and field practitioners may sometimes wish to avoid the multiplicity problem altogether by using procedural and statistical methodologies that do not involve multiple comparisons. Multiplicity effects can also be mitigated through omnibus analysis methods such as family of ANOVA´s. They can also be reduced through the use of statistical/mathematical corrections such as Bonferroni and Sidak corrections (Abdi, 2007; Nelson, 2015; Sidak, 1967) that can be applied either to desired alpha thresholds or to computed statistical results. Most importantly, analytic results from scientific research and from field polygraph testing are known to involve multiplicity effects – a distortion of analytic precision – whenever multiple statistical comparisons are used to make conclusions about the results from a scientific test or experiment.

There is reason to hypothesize that analytic results from a series of single-issue polygraph exams – the SSI hypothesis – may also be subject to multiplicity effects, like other multiple-issue polygraph exams. Previous research by Barland, Honts and Barger (1989) supports this possibility, and the need for this project. One difference between the SSI hypothesis – that precision or decision accuracy will be improved – and the COVID-19 context is that while the case criterion state for the COVID-19 context is uniform for the series of tests within a case, the SSI approach in the polygraph screening context involves independent target issues within each case. Raskin *et al.* (1988) reported that overall test accuracy may be optimized by using questions with uniform criterion states.[6]

This project employs simple and common statistical and analytic methods to examine the effect sizes for a series of single-issue screening polygraph tests using extant knowledge about the accuracy of single-issue test formats currently used in contemporary polygraph field practice.

[6] In field polygraph practice, the use of questions with uniform criterion states will mean that an examinee is either lying to all or truthful to all of the investigation target questions.

Method

Test sensitivity, specificity and error rates for single-issue polygraph examinations were obtained from a meta-analytic survey of validated polygraph techniques (APA, 2011). The aggregated test sensitivity for multiple issue test formats included in the meta-analytic survey was .771, and test specificity was .719 for these formats. The false-negative and false positive rates for multiple-issue polygraph techniques was reported as .113 and .144, respectively. For single-issue techniques aggregated test sensitivity was .741, and specificity was .702. The aggregated false-negative rate for single-issue techniques was .062, and the aggregated false-positive rate for these techniques was reported as .091. The percentage of correct decisions was reported as .904 for single-issue techniques and .850 for multiple-issue polygraph techniques. These descriptive metrics are shown in Appendix A, and were used as seed values for a simple Monte Carlo analysis of the accuracy effect sizes for SSI exams, when each case consists of a series of four single-issue polygraph tests.

The analysis consisted of 1000 iterations of a Monte Carlo space of n =1000 cases, each of which consisted of a series of four single-issue exams[7] used to investigate four different behavioral targets. For each iteration of the Monte Carlo space, a single classification [positive, negative, inconclusive] was made for each of the SSI exams within each case. The overall result of each case was coded for true-positive, true-negative, false-negative, false positive, and inconclusive outcomes based on the series of four test results.

A classification of each case was made based on the results of the SSI exams. Cases were classified as truthful if the results for all single-issue targets were truthful and were classified as deceptive if the results for one or more single-issue targets were deceptive. This all-or-any classification scheme is allegorical to the decision rule used by field polygraph examiners who conduct multiple-issue screening exams and permits a direct and intuitive comparison of effect sizes for multiple-issue exams and the SSI hypothesis. [See Nelson (2018) for a discussion of polygraph decision rules.]

Because each of the SSI exams address different behavioral target issues, for which it is conceivable that a person may engage in none, some, or all of them, the criterion state for each behavioral target was independent within each case. The criterion state of each case was innocent[8] when the criterion state of all target behaviors was truthful, and the case criterion state was guilty if the criterion state of one or more targets was guilty. The experiment-wide prior base rate for guilt was set at .5. This was done by setting the base rate at 1-(1-.5) ^ (1/4) = .159 for each individual target. For each question in each case in each iteration of the Monte Carlo space, a random number was compared to this prior and the question criterion state was set to innocent if it exceeded the prior and guilty if it did not. The case criterion state was set to guilty if any question criterion was guilty and was set to innocent if all question criterion states were innocent. The result of this process was that the criterion state was innocent for all target issues – all SSI exams for the series of exams within each case – for approximately half of the cases in the Monte Carlo space. With numerous iterations of the Monte Carlo space, the mean incidence rate for guilty cases converged to .5.

Some polygraph data analysis methods employ a statistical correction for multiplicity effects that occur when making truthful case

[7] O'Burke (2022) refers to the proposed polygraph testing approach as the single-issue-screening-test (SIST), though this term is potentially confusing because each case actually consists of a multiplicity of four target issues. O'Burke attempts to capitalize on the precision of single-issue diagnostic polygraph techniques, which may be reasonable when considering that the precision of the technique stems from the single-issue design for which more data is available to support a single result. However, O'Burke overlooks that the serial-single-issue approach remains probabilistic and is still subject to cumulative error effects under a requirement for multiple probabilistic results to achieve an overall case classification.
[8] In this usage the terms "innocent" and "guilty" do not refer to a legal judgment but are used to denote the actual criterion state in a way that is distinct from, and less easily conflated with the "deceptive" or "truthful" classification of test results.

classifications with multiple-issue screening polygraphs. The effect of this is to reduce the rate of inconclusive results for innocent cases, in addition to preventing a reduction of test specificity that results from a requirement for negative results for all investigation targets in order to classify a case as negative.[9] However, aggregated effect sizes reported by APA (2011) do not separate results with and without a statistical correction. O'Burke (2022) did not describe any use of a statistical correction for the SSI approach, and for this reason no statistical correction was used when classifying the Monte Carlo cases.

In addition, to studying the effect sizes for SSI exams with independent criterion states, accuracy effects were also studied for serial polygraph testing in which the criterion state was uniform for the series of exams within each case. This condition more closely resembles the use of serial testing strategies in other contexts (e.g., serial testing for COVID-19).

Analysis

A number of accuracy effect sizes were computed for the SSI hypothesis using a Monte Carlo bootstrap procedure and are shown in Table 1. These include test sensitivity and specificity rates, in addition to the false-negative and false-positive error rates. Standard errors were computed and the upper and lower limits of the 90% confidence interval (CI). [Refer to Nelson (2020) for a description of test accuracy metrics.] All computations were completed using the R statistical computing language (R Core Team, 2022).

Table 1. Test accuracy metrics

Metric.	Meaning
Sensitivity (TP)	True positive rate (guilty cases that are classified correctly as deceptive)
Specificity (TN)	True negative rate (innocent cases that are classified correctly as truthful)
False negative (FN)	False positive rate (guilty cases that are classified incorrectly as truthful)
False positive (FP)	False negative rate (innocent cases that are classified incorrectly as deceptive)
Guilty inconclusive (G-INC)	Guilty inconclusive rate
Innocent- inconclusive (I-INC)	Innocent inconclusive rate
Unweighted inconclusive (INC)	Unweighted inconclusive rate = (G-INC + I-INC) / 2
Guilty percent correct (GPC)	Proportion of correct guilty cases without inconclusive results = TP / (TP + FN)
Innocent percent correct (IPC)	Proportion of correct innocent case without inconclusive results = TN / (TN + FP)
Unweighted accuracy (ACCY)	Average of guilty percent correct and innocent percent correct = (GPC + IPC) / 2
Negative predictive value (NPV)	Ratio of TN and all negative cases = TN / (TN + FN)
Positive predictive value (PPV)	Ratio of TP and all positive cases = TP / (TP + FP)

Results

Two versions of the Monte Carlo experiment were analyzed. The first experiment involved the SSI hypothesis with four independent criterion states. This condition resembles the screening context described by O'Burke (2022) involving polygraph screening questions about an examinee's possible involvement in illegal drug use, serious crimes, domestic violence, and sex crimes. For field polygraph examiners, it is conceivable that an examinee may engage in none, some, or all of these activities. In practice, multiple-issue polygraph test results are interpreted with an assumption that the criterion states vary independently for differ-

[9] A statistical correction is not used for deceptive classifications with multiple-issue screening polygraphs because doing so may reduce screening sensitivity. In contrast, a statistical correction is used when making deceptive classifications based on the sub-total scores of single-issue polygraphs, in order to reduce the inflation of false-positive errors. For single-polygraphs test sensitivity is maintained by the use of the grand-total score, for which no correction is used.

ent target issues. This means that results are made first at the level of each target issue, and then parsed to achieve an overall test result. The procedural heuristic is that the overall case result is classified as truthful when the result for all target issues is classified as truthful, and the overall case result is classified as deceptive when the result for any target issue is classified as deceptive.[10]

The second experiment involved the SSI hypothesis with uniform criterion states. This condition more closely resembles a process of repeated testing of the same investigation target. This is more similar to the serial testing strategy with COVID-19, with the important difference that the goal here is to observe the effects after a series of four tests.

SSI investigation of multiple independent target behaviors

Results of the Monte Carlo analysis and bootstrap are shown later in Table 2. Application of SSI exams to four independent targets within each case produced a false-negative error rate of <.001 and a false-positive error rate was .371. The inconclusive rate for guilty cases was .126 and was loaded for innocent cases with a rate of .270.[11] The SSI approach to four independent targets produced a high-test sensitivity rate – the proportion of guilty cases that are correctly classified – of .874 with a 90% CI from .851 to .897. Test specificity – the proportion of innocent cases that are correctly classified, was lower, with a mean of .350 and 90% CI from .326 to .394.

Unweighted accuracy – the unweighted mean of the percent correct for guilty and innocent cases, excluding inconclusive results – was .746, with a standard error of .013, and a 90% CI from .725 to .768 for the SSI approach with four independent target issues. The unweighted inconclusive rate was .198. The upper limit of test specificity did not exceed the .5 level for SSI exams with either mixed criterion states.

SSI investigation of uniform target behaviors

To provide further information and insight into the effect sizes of a serial testing strategy, a second Monte Carlo analysis was completed, involving the use of SSI exams for which the test target issues, and therefore the criterion states, were uniform for a series of four exams for each case. That is, a single criterion state was set for each case, and all tests within each case had the same criterion state. The prior base rate of guilty cases was set to achieve a mean of .5 for all iterations of the Monte Carlo model. Analytic results for serial single-issue exams with uniform test targets are included in Table 2. This condition more closely resembles a process of repeated testing of the same investigation target.

Application of the SSI approach for which the target issue and criterion state were uniform for all investigation targets within each case produced an unweighted accuracy rate – the unweighted mean of the percent correct for guilty and innocent cases, excluding inconclusive results – of .747 (.726 to .768), with an unweighted inconclusive rate of .135 (.119 to .152). Inconclusive results were loaded for innocent cases, with a rate of .001 for guilty cases and. .269 for innocent cases. Classification errors were also loaded for innocent cases, with a rate of <.001 for false-negative errors, and .370 for false-positive errors. Test sensitivity to deception increased to .998 for the series of four exams with a uniform target issue. Test specificity was reduced to .359. Interestingly, the upper limit of the confidence interval for SSI exams with uniform criterion states did not exceed the .5 level.

Discussion

Validity of the SSI hypothesis would be indicated by some improvement over the effect

[10] Similar to omnibus methods in other areas of data analysis, interpretation of negative (i.e., truthful) results is not warranted when statistically significant or positive (i.e., deceptive) results are observed anywhere within an omnibus test or experiment. In scientific data analysis, further, post-hoc, analysis when there is interest or need for more information about the locus of a statistically significant result. [See Nelson (2018) for more information on polygraph decision rules.]

[11] Inconclusive rates in field practice are known to be lower than reported in published studies. This is because field polygraph examiners are permitted to engage in standardized and evidence-based methods to resolve inconclusive results, whereas information in scientific studies is more likely reported without attempting to manipulate or resolve inconclusive results.

sizes for multiple-issue polygraph exams. Table 2 shows a summary of the results of this study using a series of single-issue exams with both uniform and (non-uniform) criterion states. Also shown in Table 2 is a summary of accuracy effects for multiple-issue exams (shown in Appendix A). SSI exams uniform criterion states are exemplary of repeated testing situations in which multiple test results are desired for the same investigation target. SSI exams with mixed criterion states are exemplary of the use of a series of single-issue screening polygraph exams instead of a multiple-issue polygraph screening exam.

Table 2. Accuracy summary for serial single-issue exams with independent (mixed) criterion states, including Mean (SE) [90% CI].

Metric	4 SSI Exams (independent criterion states)	4 SSI Exams (uniform criterion states)	Criterion Independent (multi-issue) PDD Techniques†
Sensitivity (TP)	.874 (.015) [.851 to .897]	.998 (.006) [.978 to >.999]	.771 (.072) [.630 to .911]
Specificity (TN)	.359 (.022) [.326 to .394]	.359 (.065) [.255 to .477]	.719 (.047) [.626 to .811]
False negative (FN)	<.001 (<.001) [min=<.001, max=<.001] *	<.001 (<.001) [min = <.001, max=.004] *	.113 (.058) [.001 to .226]
False positive (FP)	.371 (.022) [.336 to .404]	.370 (.022) [.334 to .404]	.144 (.039) [.066 to .221]
Guilty inconclusive (G-INC)	.126 (.015) [.103 to .149]	.001 (.001) [min=<.001, max=.008] *	.112 (.051) [.013 to .212]
Innocent- inconclusive (I-INC)	.270 (020) [.235 to .304]	.269 (020) [.235 to .302]	.136 (.031) [.076 to .196]
Unweighted inconclusive (INC)	.198 (.013) [.177 to .218]	.135 (.010) [.119 to .152]	.125 (.029) [.068 to .183]
Guilty percent correct (GPC)	.999 (<.001) [min=>.999, max=>.999] *	>.999 (<.001) [min=.996, max=>.999] *	.873 (.066) [.744 to .999]
Innocent percent correct (IPC)	.492 (.026) [.450 to .536]	.493 (.026) [.452 to .536]	.831 (.043) [.746 to .915]
Unweighted accuracy (ACCY)	.746 (.013) [.725 to .768]	.747 (.013) [.726 to .768]	.850 (.039) [.773 to .926]
Positive predictive value (PPV)	.705 (.013) [.685 to .725]	.732 (.011) [.714 to .751]	.828 (.059) [.712 to .943]
Negative predictive value (NPV)	>.999 (<.001) [min=>.999, max=>.999] *	>.999 (.001) [min=>.999, max=>.999] *	.878 (.049) [.782 to .973]

* Min and max observed values are shown because observed values are so close to the limits that data are not normally distributed, making quantile functions somewhat unstable and uninformative in this context.
† Previously shown in Table 1. Added here to facilitate comparison of the means and CIs.

Effect sizes for multiple-issue polygraph techniques, described by APA (2011) as those interpreted with an assumption of independent criterion variance, are shown in Appendix A and include an unweighted accuracy rate of .850 with an inconclusive rate of .125.[12] Test sensitivity for multiple-issue exams is shown as .771, and sensitivity as .719. Inconclusive rates for multiple issue exams are reported as .112 for guilty cases and .136 for innocent

[12] As with the APA (2011) report, multiple issue exams for this analysis are characterized by pragmatic assumptions that the criterion states may vary independently for the target issues within each exam. This assumption is manifested in the selection of decision rules used to parse the categorical case result from the test data. Differences in effect sizes have not been demonstrated for different target behaviors, and polygraph validation studies have not been published for different operational target issues.

cases. The false-negative error rate for multiple-issue polygraphs was reported as .113 and the false-positive rate was reported as .144.

Results from the SSI experiments showed an increase in test sensitivity for SSI exams with uniform criterion states, along with a decrease in false-negative errors for SSI exams with both uniform and mixed criterion. The greatest observed difference was a large decrease in test specificity for the SSI approach. The mean accuracy rate was .746 for SSI exams with mixed criterion states, and .747 for SSI exams with uniform criterion states, which was lower than the reported mean accuracy for polygraph exams formulated with questions intended to be interpreted with an assumption that the criterion states of the RQs vary independently.

All the previously reported confidence intervals shown in Appendix A, for single-issue and multiple-issue exams, have substantial overlap indicating that reported differences have not been statistically significant. The meaning of this is that attempts to assert that multiple issue exams are inherently less accurate than single-issue exams are not supported by published scientific evidence. However, the means or locations (point estimates) of the distributions of effect sizes for multiple-issue exams sometimes appear to be weaker than for single-issue exams. A variety of factors may play a role in the observed differences. These may include psychological factors such as divided attention during testing, in addition to statistical multiplicity effects that compound the rates of errors and inclusive results when case classifications are based on multiple statistical calculations within each case. Another interesting aspect of multiple-issue screening polygraph is that different target issues may have different priors. In contrast, diagnostic tests, in response to a known problem, allegation, or incident, are characterized by questions that are non-independent and for which the priors are more easily characterized as uniform.

Analogies are sometimes helpful to illustrate an abstract point. Analogies from another context are sometimes especially helpful because they can help to create a more abstract and less myopic perspective. Consider the context of weapons (which sometimes invokes the same kinds practical and pragmatic compromises as science and technology). Which are better: rifles or shotguns? Intuitively we know this is not as simple as it may at first seem. It may be tempting to try to suggest that rifles are more precise and therefore better. However, if *better* is defined as more accurate, and if *accurate* is defined as the ability to hit desired targets, then we can think of situations in which shotguns are sometimes more effective at hitting a target, and therefore better. We can also think of situations in which rifles may be more effective. The point here is that notion of better is often a matter of context.

Shotguns, for this discussion, can be thought of as analogous to an omnibus analysis in scientific research, or a multiple-issue test in field polygraph testing, in which everything is evaluated at once. There are advantages to this if we expect a lot of uncontrolled variances, and if there are inherent challenges to retesting e*n vivo*. Shotguns may not be an ideal solution when there is a single target issue, and a need for sufficient precision in order to avoid the observation of unintended effects. Rifles, on the other hand, may be a preferred solution when a single target is identified and when we are reasonably confident, we can achieve a level of experimental control sufficient to reduce random variation to acceptably low levels. Although not impossible, there may be inherent complications in using a single-issue solution (e.g., a rifle) with multiple targets. For example, time-constraints, practice effects, observer effects, and other uncontrolled effects may all begin to play a role in outcomes after the onset of activities.

In human interaction and performance testing, there are few tests, and few activities in general, that can be repeated without experiencing some form of re-test or practice effect. And while metaphors are useful for introducing and socializing new abstract constructs, it will be important to avoid taking any metaphor too literally. The point of this metaphor is to create an improved contextual understanding of the multiplicity issues in scientific testing. The practical meaning of this is that suggestions for the use of a single type of solution in all contexts are not supported by scientific evidence.

Multiple-issue screening polygraphs are thought to have the advantage of practical,

and economic efficiency. They also have the advantage of increased screening sensitivity to a broader range of screening issues. Increased sensitivity will correspond to a decreased false-negative error rate. The main disadvantage of multiple-issue screening polygraphs is the inability to reliably pinpoint the exact location or cause when a statistically significant result is observed. A characteristic of screening tests is that they are sometimes intended to slightly over-predict problems. Screening tests are not, of themselves, intended to make a diagnosis. Unless carefully designed so as to manage or avoid multiplicity effects, multiple-issue screening tests may increase false-positive errors to unexpected levels. None of this is unique to the polygraph context and is well-known to all researchers who are familiar with research methods involving multivariate analysis.

The main advantage of single-issue polygraphs is increased test specificity, compared to multiple issue exams. Diagnostic results, to be useful, require two factors: 1) sufficient sensitivity to the test target issue that decisions based on test results will be significantly greater than decisions without test results, and 2) specificity – an ability to exclude cases that do not exemplify the specific issue of concern – rates that are significantly greater than what can be achieved without the test result. Test sensitivity refers to the ability of a test to notice or observe the phenomena of interest when it is present. Test specificity refers to the ability to isolate the issue of interest and exclude cases that do not express or exemplify the specific issue of concern.

An ideal test would have very high metrics for both test sensitivity and test specificity, along with very low inconclusive and error rates. But most often, our scientific tests are not ideal. And because they are inherently probabilistic, scientific tests are not expected to be infallible. Scientific tests are required only to provide known and realistic estimates of sensitivity, specificity, error, and inconclusive rates. In general, scientific tests are useful if they improve our decision-making in some way that is of practical value.

Our knowledge of test sensitivity, specificity and error rates is what permits us to make informed and strategic use of test results, despite known limitations. Rational and judicious use of test results is sometimes more intuitive when sensitivity, specificity and error rates are balanced. Conversely, strategic use of scientific test results can be less intuitive, and more difficult, when these are imbalanced. This difficulty can be mitigated by understanding the test accuracy characteristics, including the source of observed imbalance. One common source of imbalanced test accuracy is the problem of multiple comparisons – multiplicity effects.

In simple terms, multiplicity effects reduce the precision of statistical estimations and categorical conclusions. This reduction in overall precision occurs as a function of increasing the opportunity for error variance. Correct classifications can be thought of as influenced primarily by diagnostic variation. Classification problems can be thought of as overly influenced by error variance. Error variance can take several forms, including systematic error and uncontrolled or unexplained error. Systematic error variance, sometimes referred to as *bias*, is influenced by imperfection or misunderstanding in the available knowledge or information that is used to develop a scientific test or experiment. Uncontrolled error is sometimes referred to as *random error* or unexplained error.

Precision or accuracy of test results or experimental results is achieved by understanding and reducing the potential sources of error to the extent possible. Testing and research methods that involve classifications or estimations based on multiple statistical comparisons will inevitably include more opportunity for errors than those that require a single statistical comparison. However, many research and testing contexts are faced with the need to evaluate multiple questions or hypotheses, which necessitates that professionals who are involved in scientific testing and research are adequately informed about multiplicity effects.

Efforts to mitigate the effects of statistical multiplicity can be thought of as falling into three main areas. The first of these is to avoid decisions, estimations, and classifications based on multiple statistical comparisons whenever possible. This is not always possible in the polygraph screening context, whenever an agency has identified multiple target

behaviors of interest to screening decisions. As a second strategy to mitigate the effects of multiplicity in scientific testing and research is to make use of statistical corrections to alpha boundaries and decision cut-points. Some polygraph data analysis methods do make use of statistical corrections to reduce the effects of statistical multiplicity. A third approach to mitigating the effects of statistical multiplicity, and the subsequent reduction of precision for analytic results, is to make use of omnibus analysis methods such as the family of ANOVAs that can evaluate multiple testing or research hypotheses in a single analytic procedure.

Analytic results

The SSI approach resulted in a reduction of FN errors to a very low level, Inspection of the statistical confidence interval indicates the observed difference is significant. However, there is a corresponding increase in FP errors (mean = .371, 90% CI .336 to .404) along with an increase in inconclusive results for innocent persons (mean = .270, 90% CI .235 to .304). Inspection of the confidence intervals also indicates that the reductions observed in test specificity, from .719 to .359 (99% CI .326 to .394), and unweighted accuracy, from .850 to .746 (90% CI .725 to .768) were also significant.

Results from the second analysis using a series of four examinations with uniform target issues revealed a similar pattern of effects. Test specificity is significantly reduced, and FP errors are significantly increased, along with a significant increase in inconclusive results among innocent persons. The significant increase in test sensitivity is even greater when criterion states are uniform (compared to the SSI approach with mixed criterion states, shown in Table 3). Unweighted accuracy (.747, SE=.013, 90% CI .726 to .768) for a series of single-issue polygraphs with uniform criterion states was significantly lower than reported accuracy rates for a single event-specific (single-issue) examination (shown in Appendix A).

In the multiple-issue polygraph screening context, statistical multiplicity effects are present regardless of whether multiple issues are investigated using a multiple-issue test format or the SSI approach.

For practical purposes, serial, or repeated testing of uniform target issues, can be thought of as a matter of retest reliability. Most polygraph reliability studies in the past have not addressed re-test reliability and have instead addressed questions of inter-rater (inter-scorer) reliability. An exception to this trend can be seen in the analog by Honts, et al., (2015) in which repeated testing effects were not observed to reduce the availability of recorded data in the comparison question test. Increased use of automated analysis algorithms can substantially mitigate most remaining concerns about inter-scorer reliability. And though this analysis was not designed to address the question of retest reliability, these results suggest a need to better understand repeat testing and retest reliability in both polygraph screening and diagnostic contexts. One important difference between this project, involving a series of four examinations, is that the more common re-examination context may involve a single repetition of an examination.

Ancillary analyses of serial testing strategies

To better understand the retesting effects associated with repeated testing strategies with a single target issue, in contrast to the SSI hypothesis applied to multiple target issues, two ancillary analyses were completed. The first ancillary analysis involved a computation of the potential effects associated with a recommendation or requirement for repeated negative test results during the COVID-19 pandemic. This type of requirement may be associated with procedures for release from quarantine or discharge from hospital care. A second ancillary analysis involves the computation of sensitivity, specificity and error rates for re-examination of a single-issue diagnostic polygraph test.

Serial testing for COVID-19

The recent COVID-19 pandemic provides an interesting point of juxtaposition to illustrate and understand the potential use of serial testing strategies more fully. The United States government (Food and Drug Administration, 2022a) extended an emergency use

authorization and further authorized the use of point-of-care rapid-antigen (RA) tests for serial testing of COVID-19 with both symptomatic and asymptomatic persons. Additional guidance was provided a short time later (Food and Drug Administration, 2022b) in the form of a recommendation for serial testing – a second COVID-19 RA test, one or two days later – following a negative test result in order to reduce the risk of false-negative error.

Meta-analytic results (Dinnes et al., 2022) showed that RA tests for COVID-19, such as those approved for use at home, at point of care sites such as when a patient is quarantined, and for travel approval, have a sensitivity rate of .73 with persons who are symptomatic, and .55 with persons who are asymptomatic. Specificity rates for these tests were reported at >.99 for symptomatic persons and >.99 for asymptomatic persons. The corresponding false-negative error rate was .27, and the corresponding false-positive error rate was <.01.

The base-rate or prevalence of COVID-19 was also a consideration. COVID-19 was considered a public health hazard because of the high rate of transmission together with the high potential for serious health complication including death. Despite these concerns, a majority of people in the community will not have COVID-19 at any given time. For this project, a suitable base-rate would need to be determined.

Doernberg *et al.* (2022) surveyed healthcare workers in San Fransisco during the early months of the pandemic, from May to September 2020. Healthcare workers, who were considered to have a higher-than-average rate of exposure to the virus, were found to have a baseline prevalence of approximately 1% during that time. Another, equally important consideration is that despite the pandemic nature of COVID-19, most persons in will not be positive for the virus. Kalish et al. (2021) reported an analysis of data from the first six months of the pandemic, using quota sampling of over 9000 persons from a volunteer pool of over 460,000 persons, and showed a rate of approximately five undiagnosed COVID-19 cases for every diagnosed case. An important practical consideration is that prevalence rates or base rates – also referred to as incidents rates – are always a function of time at risk. Higher rates tend to be observed over longer periods of time. Ignoring the time-at-risk aspect for simplicity, the prior base rate for this ancillary analysis was set at .05 (5%) based on available published information.

Accuracy metrics, along with the estimated prevalence rate, were used as input parameters for the Monte Carlo model that was used to further compute the accuracy of single and serial COVID-19 RA tests, including the expected confidence range for PPV, NPV and overall test accuracy. The Monte Carlo space consisted of 1000 iterations of n=1000 cases. Each case consisted of two RA tests, for which the input sensitivity, specificity, and error parameters were obtained from the meta-analysis by Dinnes et al. (2022).

Cases in the serial COVID-19 RA analysis were classified as positive if either of two tests gave a positive result. COVID-19 RA test results that are not positive are classified as negative [13], with no inconclusive zone. Cases were classified as negative when both test results were negative. The effect of this decision rule was to increase both test sensitivity and test specificity. Results are shown in Table 3.

[13] In contrast, comparison question polygraph tests are evaluated for the level of statistical significance for both deception and truth-telling. It is possible that some polygraph results are not statistically significant for either – for which the contextual categorical terms inconclusive and no-opinion are used by field polygraph examiners.

Table 3. Accuracy estimation for COVID-19 rapid-antigen tests with prior = 5%, used in both single and serial testing contexts, including the mean (SE) and [90% CI].

	Single COVID-19 RA test	Serial COVID-19 RA test
Unweighted accuracy†	.860 (.030) [.810 to .909]	.953 (.018) [.922 to .980]
Sensitivity	.730 (.060) [.630 to .833]	.925 (.035) [.880 to .981]
Specificity	.990 (.03) [.984 to .995]	.980(.005) [.973 to .987]
FN Errors	.270 (.06) [.167 to .370]	.075 (.035) [.019 to .130]
FP Errors	.010 (.030) [min=.001, max=.024] *	.020 (.005) [.013 to .027]
PPV	.809 (.054) [.720 to .898]	.730 (.047) [.654 to .810]
NPV	.985 (.003) [.979 to .990]	.996 (.002) [.992 to .999]

† Unweighted accuracy can be thought of as the accuracy estimate if the seropositive and seronegative group sizes were balanced. In this analysis the actual percent correct is less informative because 95% of the cases were innocent.
* Min and max observed values are shown because observed values are so close to the limits that data are not normally distributed, making quantile functions somewhat unstable and uninformative in this context.

Not surprisingly, NPV was high (.985) for a single COVID-19 RA test, because of the high-test specificity combined with the high prior probability of being negative and remained high for the serial testing situation. However, the FN error rate for a single RA test was .27, which may be excessive for some risk management purposes. The goal of the serial testing strategy was to decrease FN errors. Table 5 shows that the FN error rate was reduced to .075 by administering a second RA test. The increase in FP errors from .01 to .02, with the corresponding decrease in PPV, from .809 to .730 is not surprising and may be of little concern in this context. However, other testing contexts may be more concerned about these effects and may therefore require a different testing strategy.

Serial testing in the diagnostic polygraph context

Returning to the polygraph test, serial diagnostic testing (SDT) strategies have been observed in the context of re-examination or repeating an event-specific diagnostic polygraph test. Table 4 shows the Monte Carlo results of SDT of a single-issue. The Monte Carlo model consisted of 1000 iterations of n =1000 single-issue cases. Each case consisted of two exams for which the guilty or innocent criterion state was identical for the two exams. The input parameters were obtained from the information in Table 1, for single-issue polygraph exams.

The SDT approach differs from the successive hurdles screening (SHS) strategy (Krapohl & Stern, 2003). SHS involves the serial use of multiple-issue screening tests followed by the use of more precise single-issue screening tests before classification of a case as positive. [14] In the present SDT analysis cases were classified as deceptive when the results of both tests were deceptive and were classified as truthful when the results of both tests were truthful. This classification scheme is different from that of the SSI hypothesis in which

a single positive result would be used to make deceptive classifications. SDT can be used with high-interest or high-values cases – such as those intended for use in a legal proceeding – and involves the re-examination of a positive or negative single-issue test result.

Table 4. Accuracy summary for repeated concordant event-specific polygraphs.

Metric	Mean (SE) [90% CI]
Sensitivity (TP)	.704 (.021) [.668 to .738]
Specificity (TN)	.600 (.023) [.563 to .638]
False negative (FN)	.006 (.003) [min = <.001, max=.020] *
False positive (FP)	.012 (.005) [.004 to .020]
Guilty inconclusive (G-INC)	.291 (.021) [.258 to .326]
Innocent- inconclusive (I-INC)	.388 (022) [.350 to 423]
Unweighted inconclusive (INC)	.339 (.015) [.314 to .364]
Guilty percent correct (GPC)	.992 (.005) [.983 to >.999]
Innocent percent correct (IPC)	.981 (.008) [.966 to .993]
Unweighted accuracy (ACCY)	.986 (.005) [.978 to .994]
Positive predictive value (PPV)	.984 (.007) [.972 to .994]
Negative predictive value (NPV)	.991 (.006) [.981 to >.999]

* Min and max observed values are shown because observed values are so close to the limits that data are not normally distributed, making quantile functions somewhat unstable and uninformative in this context.

Results shown in Table 4 illustrate the potential value of SDT and the use of concordant test results. Overall accuracy increased to .986 with PPV=.984 and NPV=.991, along with correspond reduction of both FN and FP errors to very low levels. However, the effective sensitivity and specificity rates for SDT were lower than for single examinations. This was related to a substantial increase in inconclusive results – over one-third of the cases – when results of the two tests did not concur. [Refer to Nelson (2016) and Nelson and Turner (2017) for information and discussion about the effects of serial testing in which concordant

[14] SHS is analogous to the medical use of screening and diagnostic tests, and analogous to the use of omnibus and post-hoc tests in scientific research.

results are not required – in which one test result is selected as a basis for the classification of deception or truth telling.[15],[16]]

SDT strategies have been described previously (Handler, 2016; Kircher & Raskin, 2017. The potential advantages of SDT were first described by Meehl and Rosen (1955). Calculation of the accuracy of two test results involves the use of the phi correlation coefficient, and was described by Nelson, Kircher and Handler (2018). An important aspect of SDT is that it is critical the two tests are independent. That is, there must be not source of shared variance or contamination between the two tests. In other words, independence means that the results of one test cannot possibly influence the results of another test. In practice this may be difficult to achieve if an examiner at the second test is informed about the results of the first test. Another potential source of errors is the possibility that there are inherent vulnerabilities related to the test methodology or technology. For this reason, serial testing in the medical context will often involve the use of two tests that make use of different data and different methods of analysis. In the polygraph context this might involve the serial use of two different credibility assessment test technologies.

Conclusion

Overall, these results are consistent with previously reported findings (Barland *et. al.*, 1989) that failed to support the SSI hypothesis as advantageous when compared with the results of traditional multiple-issue screening polygraphs. Although not discussed in terms of statistical phenomena at the time, the results of Barland *et al.* are easily understood today as a matter of statistical multiplicity or cumulative statistical error. The SSI hypothesis amounts to the use of multiple statistical classifications, albeit in a slightly different way than multiple issue polygraphs and is subject to similar distortions and potential reductions of analytic precision.

The SSI approach to multiple-issue polygraph screening is allegorical to proceeding immediately to the post-hoc analysis in scientific research – without first conducting an omnibus analysis. The effect will be to increase family-wise (experiment-wise) error rates. In the polygraph context the increase in errors and inconclusive results are loaded on innocent cases. Omnibus analytic solutions have the effect of reducing experiment-wise error rates. In the polygraph context this can provide an improved balance of test sensitivity and specificity rates.

Evidence currently does not support that the SSI hypothesis provides a simple and convenient solution to the multiple-issue polygraph screening context. Overall accuracy may be inadvertently reduced by this approach, with corresponding increases in errors and inconclusive results that are loaded on innocent examinees. However, this does not imply that the SSI hypothesis does not provide some potential value to polygraph programs that correctly understand multiplicity and other effects discussed above. For those interested in strategically increasing test sensitivity while reducing FN errors, and who are correctly informed about the reduction of test specificity and accuracy, the potential value of the SSI approach may be determined by economic considerations in addition to test accuracy metrics.

An obvious limitation of this project is that it did not investigate the potential use of statistical corrections to reduce multiplicity effects with either the traditional multiple-issue polygraph screening

Another obvious limitation of this project is that it did not investigate the potential use of statistical corrections to reduce multiplicity effects with either the traditional multiple-issue polygraph screening test or the SSI approach. This is in part because of the design of the Monte Carlo model, using seeding parameters in the form of test sensi-

[15] In general, classification of a test as deceptive or truthful based on a single test result from a serial testing context may increase the likelihood of error.

[16] Nelson (2016) and Nelson and Turner (2017) also described the effects when using the output of an initial test as a Bayesian prior for a subsequent test. This was also discussed by Handler (2016) and Kircher and Raskin (2016).

tivity, specificity and error rates that are not easily subject to statistical correction. Another limitation is that, as a Monte Carlo study, it did not involve the analysis of analog or en vivo data. This is somewhat mitigated by the consistency of these results with the results of a previous analog study (Barland *et. al.*, 1989), that compared multiple-issue polygraph test results with a series of single-issue exams.

A more important limitation of this project is that it did not investigate the variety of economic factors that must be taken into consideration by program managers and policy makers who may be considering the SSI approach to polygraph screening. Economic factors include both fiscal economics and physical economics. While fiscal economic factors relate to the direct costs of completing multiple single-issue polygraphs instead of a single multiple-issue exam, physical economic factors can include increased testing time, and scheduling for professionals, facilities and instrumentation. Economic factors also include increased quality control requirements that are necessary to ensure standards compliance with a series of four examinations. Social and psychological economic factors may include things such as increased testing and interview fatigue when setting up numerous sets of RQs and CQs, increased opportunity for procedural error, habituation effects, practice effects, and observer effects, along with increased opportunity for behavioral noncooperation.

Economic costs can include both short-term and long-term functions involving test outcomes. It may be that short terms costs are more immediately obvious to agencies and field polygraph professionals. Inconclusive results lead to repeated testing and to increased interrogation of innocent persons. They may lead also to an increased opportunity or risk of false confessions. Another obvious economic cost is the potential reduction of the available pool of suitable applicants in public safety agencies. Potential costs for FN errors can include an increased risk of harm to an agency or community, and loss of professional reputation for an examiner. Potential costs for FP errors are loaded for the examinee. However, this does not imply that FP errors are cost-free. More likely, the economic costs of FP errors to agencies and communities, and to the polygraph profession, may be observed over longer periods of time. It remains for program managers to study and understand the cost functions associated with FN and FP errors, and to implement programmatic strategies to achieve goals and reduce costs.

Yet another limitation of this project is that no effort was made to investigate accuracy effects for multiple issue examinations with mixed priors. That is, when different test target issues have different prevalence or incidence rates. It is not likely that all investigation target issues in polygraph screening programs have similar base-rates of occurrence. Related to this is the fact that statistical classification and probabilistic inference activities will be prone to a higher false-positive-index (FPI, the ratio of false-positive and all positive results) rates when the prior probability is low, while the false-negative-index (FNI, the ratio of false-negative and all negative results) will be higher when the prior probability is high. In many testing circumstances – such as when attempting to identify or diagnose problems – it is the test sensitivity and false-positive metrics that are of greatest interest. However, other testing contexts – such as those interested in ruling out a particular diagnosis or problem – may be more interested in test specificity and false-negative metrics. Of importance here is that multiplicity effects can involve both positive and negative test outcomes. Equally important, multiplicity effects can occur regardless of the prior.

A final limitation of this project is that it did not investigate the benefits, in terms of precision or economics, related to post-hoc testing practices – referred to by field polygraph examiners as break-out tests. Somewhat similar to scientific research involving omnibus analytic methods and post-hoc testing, a variety of post-hoc polygraph screening approaches have been suggested. It is possible that there is some value in each of the different suggestions, and that the selection of an optimal post-hoc approach may be a matter of economic and contextual factors.

Future research activities should be developed to further advance our knowledge of the role of statistical multiplicity, and the potential use of simple statistical corrections, in polygraph screening programs. At this time

data indicate that testing multiple independent target issues will produce more testing errors and inconclusive results than testing a single target issue – regardless of whether using traditional multiple-issue polygraph techniques or the SSI approach. More information is needed around the variety of economic factors that are taken into consideration when developing polygraph programs and polygraph field practice policies. More research is also needed around the effects of multiple-issue exams with mixed priors. More training may be needed among polygraph field practitioners around managing and correcting for multiplicity effects, and the strategic use of omnibus and specific analytic strategies. Additional training may also be needed in working with prior information, and the serial use of test results in Bayesian classification.

The present analysis is not the first attempt to develop or achieve analytic results to inform our knowledge about the SSI. Previous research (Barland *et al.* 1989) also failed to support the notion that the SSI approach to multiple-issue screening is inherently superior or advantageous. Some effects of the SSI approach may be desirable, Decreased FN errors and increased test sensitivity may be a useful advantage to high-value polygraph screening situations with a large applicant pool.[17] However, economic factors and increased FP errors, increased inconclusive results, and reduced overall precision may be problematic for other polygraph screening programs.[18] Data at this time do not support the abandonment of multiple-issue polygraph techniques. Similar to the broader scientific context, available evidence indicates that there remains important value in the strategic use of omnibus solutions. This analysis suggests that the SSI approach to multiple-issue polygraph screening is effectively a more protracted multiple-issue exam – it does not mitigate statistical multiplicity effects and remains subject to known problems associated with the use of multiple statistical comparisons.

[17] For example: screening applicants for the Uniformed Division of the U.S. Secret Service, whose officers are authorized (required) to patrol areas outside the White House while armed with automatic or semi-automatic weapons. In this situation, program managers might wish to use the SSI approach to reduce FN errors. With a large enough pool of experienced and high-quality applicants, the excess FP errors may be of lesser concern. Reduction of the alpha from .05 to .01 would also reduce FN errors, but without the corresponding increase in FP errors.

[18] For example: large metropolitan or state police agencies, that may serve as career entry points for many law enforcement professionals, may have a smaller than optimal applicant pool, and be more concerned about the long-term economic costs of excess FP errors and inconclusive results in addition to the obvious costs associated with FN errors. These agencies may prefer screening solutions with higher overall accuracy and more balanced test sensitivity and specificity rates. Smaller law enforcement agencies may also have lower applicant-to-hire ratios and may experience negative economic effects from the higher FP and inconclusive rates of the SSI approach. They may instead prefer the traditional multiple-issue polygraph approach to applicant screening, in which testing errors can be managed through the selection of alpha thresholds and statistical corrections.

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N.J. Salkind (Ed.) *Encyclopedia of Measurement and Statistics*. Sage.

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.

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

Dinnes J, Sharma P, Berhane S, van Wyk SS, Nyaaba N, Domen J, Taylor M, Cunningham J, Davenport C, Dittrich S, Emperador D, Hooft L, Leeflang MMG, McInnes MDF, Spijker R, Verbakel JY, Takwoingi Y, Taylor-Phillips S, Van den Bruel A, & Deeks JJ. (2022). Cochrane COVID-19 diagnostic test accuracy group. Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database of Systematic Reviews 2022*, Issue 7. Art. No.: CD013705. DOI: 10.1002/14651858.CD013705.pub3.

Doernberg, S. B., Holubar, M., Jain, V., Weng, Y., Lu, D., Bollyky, J. B., Sample, H., Huang, B., Craik, C. S., Desai, M., Rutherford, G. W., & Maldonado, Y. (2022). CHART Study Consortium. Incidence and Prevalence of Coronavirus Disease 2019 Within a Healthcare Worker Cohort During the First Year of the Severe Acute Respiratory Syndrome Coronavirus 2 Pandemic. *Clinical Infectious Diseases*, 75(9): 1573-1584. Doi: 10.1093/cid/ciac210. PMID: 35279023; PMCID: PMC8992269.

Food and Drug Administration (November 1, 2022a). [Letter re: Revisions Related to Serial (Repeat) Testing for the EUAs of Antigen IVDs]. Retrieved from https://www.fda.gov/media/162799/download.

Food and Drug Administration (November 17, 2022b). *At-Home COVID-19 Antigen Tests-Take Steps to Reduce Your Risk of False Negative Results: FDA Safety Communication*. https://www.fda.gov/medical-devices/safety-communications/home-covid-19-antigen-tests-take-steps-reduce-your-risk-false-negative-results-fda-safety.

Handler, M. (2016). Low base rate screening survival analysis & successive hurdles. *Police Polygraphist*, 2016(March), 31-38.

Honts, C. R., Handler, M., Shaw, P., & Gougler, M. (2015). The vasomotor response in the comparison question test. *Polygraph, 44* (1), 62-78.

Kalish, H., Klumpp-Thomas, C., Hunsberger, S., Baus, H. A., Fay, M. P., Siripong, N., Wang, J., Hicks, J., Mehalko, J., Travers, J., Drew, M., Pauly, K., Spathies, J., Ngo, T., Adusei, K. M., Croker, J. A., Li, Y., Graubard, B. I., … Czajkowski, L. (2021). Undiagnosed SARS-CoV-2 seropositivity during the first 6 months of the COVID-19 pandemic in the United States. *Science Translational Medicine*, 13 (601), eabh3826. DOI: 10.1126/scitranslmed.abh3826.

Kircher, J. C. & Raskin, D. C. (2016). Laboratory and field research on the ocular motor test. *European Polygraph, 10*(4), 159-172.

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

Meehl, P. & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

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. (2016). Five-minute science lesson: a second look at successive hurdles screening. *APA Magazine, 49*(1): 115-122.

Nelson, R. & Turner, F. (2017). Bayesian probabilities of deception and truth-telling for single and repeated polygraph examinations. *Polygraph & Forensic Credibility Assessment*, 46 (1): 53-80.

Nelson, R., Kircher, J. & Handler, M. (2018). How to calculate the expected agreement and the combined accuracy of two test results. *Polygraph & Forensic Credibility Assessment, 47*(1): 18-25.

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

Nelson. R. (2020) Five-minute science lesson: test accuracy metrics. *APA Magazine, 53*(3), 53-57.

O'Burke, J. P. (2022). Using single-issues in screening examinations. *APA Magazine, 55*(6), 51-59.

Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. J*ournal of Applied Psychology, 78*, 788-797.

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

Prado, R. Grajales, C. & Nelson, R. (2015b). Laboratory study of a diagnostic polygraph in a single sequence: a replication. *Polygraph, 44*(2), 1-12.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). *A Study of Polygraph Examination in Criminal Investigation*. Final Report to the National Institute of Justice Grant No. 85-IJ-CX-0040. University of Utah, Department of Psychology. Reprinted in Polygraph & Forensic Credibility Assessment, 48 (1): 10-39.

R Core Team (2022). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Research Division Staff (1995a). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats.* Report number DoDPI94-R-0008. DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 26 (2), 79-106.

Research Division Staff (1995b). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage*. DoDPI94-R-0009. DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 27, (3), 171-180.

Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distribution. *Journal of the American Statistical Association, 62*, 626-633.

Appendix A.

Previously reported summary of polygraph accuracy metrics

Table 1. Previously reported accuracy metrics, including the mean (SE) and [95% CI]. Taken from the APA (2011) meta-analytic survey of validated polygraph techniques Table 7 criterion accuracy for multiple issue (criterion independent) and single-issue (non-independent) polygraph techniques.

	Criterion Independent PDD Techniques	Non-independent PDD Techniques
Percent Correct	.850 (.039) [.773 to .926]	.896 (.030) [.837 to .955]
Inconclusive	.125 (.029) [.068 to .183]	.106 (.031) [.044 to .167]
Sensitivity	.771 (.072) [.630 to .911]	.840 (.050) [.743 to .938]
Specificity	.719 (.047) [.626 to .811]	.775 (.059) [.658 to .891]
FN Errors	.113 (.058) [.001 to .226]	.074 (.032) [.011 to .138]
FP Errors	.144 (.039) [.066 to .221]	.109 (.041) [.029 to .189]
D Inc	.112 (.051) [.013 to .212]	.089 (.039) [.011 to .166]
T Inc	.136 (.031) [.076 to .196]	.122 (.049) [.027 to .218]
PPV	.828 (.059) [.712 to .943]	.893 (.039) [.816 to .969]
NPV	.878 (.049) [.782 to .973]	.910 (.043) [.826 to .995]
D Correct	.873 (.066) [.744 to .999]	.919 (.036) [.849 to .989]
T Correct	.831 (.043) [.746 to .915]	.873 (.047) [.780 to .965]