Decision Accuracy for the Relevant-Irrelevant Screening Test: A Partial Replication

Donald Krapohl and Terry Rosales[1]

Abstract

In a 2005 exploratory analysis, Krapohl, Senter and Stern reported a by-case accuracy of 73% for a blind evaluator performing global analysis with a sample of field cases conducted with the Relevant-Irrelevant (RI) screening test. A total of three analytic methods had been used to evaluate the RI charts in that project, but none outperformed global analysis. However, with only a single scorer for each method the study did not answer the question of interrater reliability. Moreover, a potential confound in the ground truth criterion for one of the relevant topics was uncovered. To address these problems, we undertook a new evaluation of the RI screening test in which multiple scorers of the data were used, and included cases in which there was more confidence in the ground truth confirmation. The only method of chart interpretation was the global approach inasmuch as it represents general field practice. Overall decision accuracy was greater than chance, and averaged 87.2% for deceptive cases and 41.1% for truthful cases when No Opinion decisions were excluded, for a mean average of 64.2%. No Opinion results averaged 6.5% for deceptive cases, and 10% for truthful cases. Interrater agreement was modest but greater than chance, and low compared to other polygraph screening techniques. These findings agree with previous research suggesting the RI screening test may be valid as the first step in a multi-step polygraph screening process, though its decision accuracy may be insufficient to recommend it as a stand-alone technique.

Introduction

In 2011 a committee of the American Polygraph Association (APA) issued an exhaustive literature summary regarding the scientific evidence of various polygraph techniques. Conspicuous by its absence was any assessment of the Relevant-Irrelevant screening test (RI), a method commonly used by police and government organizations as part of the employee selection process. The stated reason for neglecting the RI was because there was only a single research study (Krapohl, Senter & Stern, 2005) that tested the RI in the manner it is used in the field. A single study was considered insufficient by the APA committee, and without replication or other verification of the findings the RI stood un-validated.

We replicated a portion of the 2005 study that involved global analysis of RI cases. We used field cases drawn from the same large sample for which there was reliable ground truth for each of the test issues. Unlike the previous research, however, multiple blind evaluators were used so that interrater agreement could be assessed.

[1] The authors are assigned to the National Center for Credibility Assessment (NCCA). The views expressed in this paper are exclusively those of the authors, and do not necessarily represent those of the NCCA, Department of Defense or US Government. Correspondence should be directed to donald.krapohl@ncca.mil.

The authors sincerely appreciate the statistical assistance and power analysis provided by Dr. Stuart Senter, and for the administrative support from Holly Cokley, Whitney Odom, and Louane Powers. A special thanks to the late Dr. Andrew Dollins for helping overcome a host of major and unanticipated technical challenges during the study. Also we appreciate the assistance of the four polygraph examiners who labored over the 100 cases, and provided their decisions for analysis. Without their participation this project would not have been possible.

The RI screening test is a multiple-issue polygraph technique in which several independent and semi-independent relevant questions are typically used to verify the background of applicants for employment. The number of relevant questions may vary, ranging from two to five. There is no published scoring method for the RI screening test, and users rely on global evaluation to assess whether the examinee has answered untruthfully, and if so, to which topic area. A more thorough summary of the background of the RI is available elsewhere (Krapohl, Senter & Stern, 2005) and is not repeated here.

The study used an archive of field cases for which there was confirmation of ground truth for each relevant question. All cases had the same four relevant questions, and used the same testing protocol. The goals of this study were to assess decision accuracy by case, and interrater agreement by case and by test question.

Method

Scorers

Four federally certified polygraph examiners with experience in conducting and evaluating the RI screening test were recruited as scorers. They had an average experience of 20.3 years as polygraph examiners (range 15-32 years), and 17.5 years with the RI screening test (range 14-25 years).

Instrumentation

The cases had been conducted on Axciton computer polygraphs (Axciton Systems, Houston, TX). The conventional four channels of data were recorded: electrodermal, relative blood pressure, and two channels of respiration. In the era in which these data were collected the use of movement sensors and photoplethysmographs were yet uncommon.

Source of Cases

All cases came from a federally funded project in the mid-1990s. A large security firm in Atlanta, Georgia, conducted RI screening cases of candidates for employment as contract security guards at the nearby international airport. There were four relevant questions in each case: 1) convictions or fines for traffic violations in the state of Georgia in the previous seven years; 2) having been granted bankruptcy in the state of Georgia in the previous seven years; 3) having used marijuana in the previous 30 days, and; 4) having ever been convicted of a felony in the state of Georgia. Confirmation of the use of marijuana was a positive urinalysis result or posttest statement from the examinee. Confirmation for traffic offenses, bankruptcy and felony convictions was based on Georgia state record checks or posttest statements of the examinee.

A total of 999 individuals applied for employment during the study period, however 217 did not appear for their scheduled polygraph examinations. One individual terminated the polygraph examination during the pre-test interview. Another case was evaluated as "no opinion" due to the examinee's inability to stay awake during the polygraph session. Technical problems resulted in the loss of data for 11 examinations. Of the 769 available cases, ground truth to all of the relevant questions was obtained on 733, with incomplete verification of the remaining 36 cases. The 733 fully confirmed cases constituted the population from which the present sample was drawn. The examinees were 97.4% African Americans, and 61.8% were female.

It should be noted that all of the paper records from this field study were destroyed as a routine matter sometime after the Krapohl, Senter and Stern (2005) report due to government space restrictions and the age of the files. Remaining were the electronic copies of the physiological data, an Excel spreadsheet showing ground truth for each case and question, and an electronic version of an incomplete draft of the researchers' original report (Brownlie, Johnson, & Knill, unpublished) from which all of the background was derived. What was not available to the present project were the decisions of the original polygraph examiners, demographic information on each examinee, and source of confirmation for each relevant question. These data gaps limited the options for subsequent analyses.

Case Selection

A power analysis was conducted to determine the samples sizes. Based on the previous research (Krapohl, Senter & Stern, 2005), we anticipated an effect size of 0.2 (70% decision accuracy over chance of 50%). Samples of 50 deceptive and 50 truthful cases allowed a power of 0.83 at an alpha of 0.05, and these are the sample sizes used in the study.

Because polygraph screening research tends to show that decision accuracy is much better in the detection of liars than of lies (correct classification of individual examinees over the correct classification of individual test questions), we considered whether the inclusion of cases with multiple deceptions might produce an inflated estimate of polygraph decision accuracy. In a reanalysis of the Krapohl et al. data (2005), correct detection of deceptive cases incrementally improved as the number of deceptions per case increased: 63% for a single deception, 84% for two deceptions, and 100% for three deceptions. This finding is consistent with expectations that more frequent deceptions will improve the performance of blind reviewers in detecting deceptive cases. However, cases with multiple deceptions were relatively infrequent among the available cases. Out of the 227 deceptive cases available, 204 (90%) were those in which the examinee was deceptive to only one relevant question.

In the field it is unknowable a priori how many issues an examinee is deceptive to, if any. We considered that at a minimum a screening technique should be able to discern deceptiveness to a single issue. Otherwise the screening method would not be useful in a majority of cases. For this reason we selected only cases in which the examinee had been deceptive to one issue, with an acknowledgement this decision may provide a lower estimate for RI validity than if the multi-deception data had been included.

To further standardize the study, we opted for keeping the deceptive issue constant across the entire sample. In an assessment of the available cases we discovered that two of the test topics did not contain a sufficient number of deceptive cases. For example, the bankruptcy topic lacked even a single confirmed deception. The topic of felony convictions held only 16 confirmed deceptive cases, significantly fewer than the minimum sample size of 50 cases. Moreover, 7 of the 16 cases had multiple deceptions, and would be excluded for reasons outlined earlier.

A third topic, conviction for traffic offenses in the previous seven years, had 90 cases in which the examinees' statements were in disagreement with ground truth. It was learned from the Krapohl et al. study (2005), however, that some unknown percentage of examinees had likely been convicted of the traffic offenses in a manner in which the examinee may not have been aware of the conviction. In many of those cases, the convictions were in absentia. Consequently, the ground truth criterion (court records) may not match the examinee's own knowledge. This created a dilemma: Should these cases be selected to take advantage of their unambiguous ground truth criterion, or should they be excluded because in some unknown proportion of the cases the examinees were unaware that they had been convicted of the traffic offenses? Ultimately we concluded that unless there could be confidence that the examinee knew ground truth, the ground truth criterion was not sufficiently reliable. Such was the case with the issue of traffic offenses. Because there was another, better alternative, the cases with deceptions to the traffic offense question were set aside.

The fourth topic, marijuana use, was seen as best for the present purposes for three reasons. First was the time window in the relevant question: only the 30 days previous to the examination. It was eminently likely that the examinees' memories would avail themselves of any marijuana experiences over such a short period. Second was the face validity of this issue. Examinees would likely grasp how a potential employer would view illegal drug use when selecting among candidates who would work in the security zones of a large international airport. Finally, there was a lot of data to choose from, with 104 cases where the examinee had been deceptive to that question alone, making it the most frequently observed deceptive topic. For these reasons we opted to randomly select from among the cases where the examinee had been confirmed deceptive only to the topic of recent marijuana use. A sample of 50 cases was also randomly drawn from the available 506 confirmed truthful cases.

Case Preparation

One concern that arose was that scorers conducting the blind evaluation of the charts may intuit a pattern in the deceptive questions, that is, over time the scorers may note the reactivity associated with the same question label over several cases. This observation may provide information that could alter their interpretation of the remaining cases. To disguise the fact that deceptions were always to the same test question, the question labels were randomly rearranged among the relevant questions on the test charts to hide which question was truly being presented. The 100 cases were then randomized, labeled 1 to 100, and the charts printed and placed in file folders in groups of 10 cases.

Data Collection

The blind scorers were informed that the purpose of the project was to gauge decision accuracy with field cases conducted with the RI screening test. They were naïve as to the base rate, test topics and ground truth of the cases. Scorers were allowed to view only 10 cases every 24 hours in order to control for fatigue.

They were instructed to evaluate each case as though they were the testing examiner. As in field practice, they were to assume that the RI screening test would be followed by more thorough interviewing and additional testing on the issue(s) on which the scorers could not conclude the examinee had been truthful. To increase the level of engagement in the effort, the scorers were also told that their individual error rates for both truthfulness and deception would be included in the report, though anonymized, so that readers could see how they compared to the other RI chart evaluators.

Imitating field procedures, the scorers were first asked to make a decision regarding the case, whether there were significant and consistent responses to any relevant question. If there were, the case would be classified SR (Significant Reactions), the scorer would then identify which question(s) they would address with post-screening interviewing or testing, and in what order. As such, the scorers would be making decisions about which steps would follow the RI screening.

Polygraph decisions were recorded and assessed for accuracy and interrater reliability. Alpha was set at .05 for all statistics, unless otherwise stated.

Results

Decision Accuracy

In field practice, at least among federal agencies that use the RI, decisions of SR are not typically made when only the screening test has been conducted. Rather, physiological responding to relevant questions prompts more interviewing and additional testing with more focused questions, a process which can ultimately lead to a decision of SR. However, for convenience and for statistical classification, when scorers reported seeing significant responding to relevant questions in the RI screening test, these tests have been deemed SR here.

Table 1 shows average decision accuracy by case. Overall decision accuracy was greater than chance when No Opinion decisions were excluded (z = 1.99, p<.05), though not when No Opinions were counted as errors (z = 1.31, ns). Most of the errors were with truthful cases, where about half of the decisions were false positives.

Table 1. Percentages of average correct, incorrect, and No Opinion (NO) decisions for 100 RI cases by four blind scorers

	Truthful Cases (n=50)	Deceptive Cases (n=50)	Overall with NO	Overall w/o NO
Correct	37.0%	81.5%	59.3%	64.2%
Error	53.0%	12.0%		
No Opinion	10.0%	6.5%		

Blind scorers correctly identified deceptive cases significantly better than they did truthful cases (z = 4.52, p<.05). While the detection of deceptive cases was greater than chance (z = 3.32, p<.05), detection of truthful cases was not (z = -1.31, ns).

In addition to making decisions by case, blind scorers were asked to list the test question(s) on which they would conduct additional testing, and place them in order from first to last. The average number of those test questions across scorers ranged from 1.0 to 2.8 with a mean of 1.7.

Table 2 lists the rank given by the four blind scorers to the correct deceptive issue for the cases in which they made correct SR decisions. In some cases the scorer made a correct decision for the case as SR, but the issues he chose for additional post-screening attention did not include the one issue to which the examinee had been deceptive.

Table 2. Number of correct SR case decisions and the scorers' rank of the deceptive issue within those cases

	Number of Correct SR Decisions by Case (n=50)	Ranking of Deceptive Issue			
		1st	2nd	3rd	4th
Scorer 1	42	33	1	0	0
Scorer 2	46	33	4	0	2
Scorer 3	37	28	0	0	0
Scorer 4	45	36	1	0	0

Interrater Reliability

By case, there were three decision options available to the blind scorers (SR, NSR, and No Opinion). Chance agreement between any two scorers was 33.3%. The observed pairwise agreement among blind scores averaged 59.7%, which was greater than chance (z = 3.74, p<.05). Most, but not all, agreement between pairs of blind scorers was above chance. See Table 3.

Table 3. Pairwise percentage of agreement for four blind scorers and ground truth for 100 RI cases

	Scorer 1	Scorer 2	Scorer 3	Scorer 4	Average
Ground Truth	60%*	62%*	75%*	40%	59.3%*
Scorer 1		73%*	60%*	55%*	
Scorer 2			64%*	63%*	
Scorer 3				43%	

* Indicates percentages statistically greater than chance.

Proportions of NSR, SR and No Opinion decisions varied substantially among the four blind scorers. The general trend for three of the four blind scorers was to make SR decisions, and there were marked differences in the rate of NSR decisions. See Table 4.

Table 4. Number of NSR, SR and No Opinion decisions for each of the four blind scorers for the 100 RI cases.

	NSR	SR	No Opinion
Scorer 1	22	66	12
Scorer 2	19	76	5
Scorer 3	51	49	0
Scorer 4	6	78	16

Table 5 covers only deceptive cases, and shows agreement between pairs of blind scorers on which relevant question each had ranked first. In some cases the blind scorers had chosen no relevant issue because he had called the case NSR. Examiners had four relevant issues to rank as first, in addition to selecting no issue, and therefore chance agreement was 20%. Pairwise agreement was significantly greater than chance for all comparisons.

Table 5. Pairwise agreement between blind scorers as to which test issue ranked first among the 50 deceptive cases.

	Scorer 2	Scorer 3	Scorer 4
Scorer 1	80%	74%	74%
Scorer 2		72%	80%
Scorer 3			70%
Average agreement			75.0%

Table 6 is similar to Table 5, except that it covers only truthful cases. Again, chance agreement was 20%. Pairwise agreement was significantly greater than chance for all comparisons. In addition, the difference between the average percentage of agreement for deceptive cases (75%) and truthful cases (53.6%) was statistically significant. This indicates that agreement on which test question required additional examiner attention post-screening was greater for deceptive cases than for truthful cases.

Table 6. Pairwise agreement between blind scorers as to which test issue ranked first among the 50 truthful cases.

	Scorer 2	Scorer 3	Scorer 4
Scorer 1	52%	54%	54%
Scorer 2		52%	61%
Scorer 3			48%
Average agreement			53.6%

Discussion

This study adds positive evidence regarding the validity of the RI screening test as a screening instrument. Overall decision accuracy was greater than chance, and as such, would support the RI screening test meeting the minimum validity requirements of the APA Standards of Practice. Consistent with earlier research, the technique is very effective in identifying deceptiveness, though significantly less so for truthfulness. This argues, as has previous work, for the use of the RI as part of a successive hurdles approach to screening. The RI as a stand-alone method incurs a significant liability owing to its weakness in detecting truthfulness, and the RI's use in this fashion is recommended here only for settings where there is a high tolerance for false positive errors.

At first blush, our finding of a mean accuracy of 64.2% (excluding No Opinions) would appear to be substantially lower than the 73.0% accuracy reported in the Krapohl et al (2005) study for the global scorer evaluating cases drawn from the same archive. This direct comparison should not be made, however, for a couple reasons. First, in the 2005 research there had been only one blind scorer performing the global analysis. It is possible that this scorer was exceptionally proficient, and did not represent the typical skill level in the field. Using group data, as we did in this partial replication, is likely to give a more stable estimate of decision accuracy in the field.

Second, recall that in the present study we restricted the sample of cases to those that were deceptive only to a single relevant question. The 2005 study included cases with multiple deceptions. This may account for some of the different results in the two studies. As reported earlier, limiting the 2005 study to only those cases in which the examinee had been deceptive to a single relevant question produced a mean accuracy of 63.0%. This is virtually identical to the present findings of 64.2%. The accuracy estimate reported in the 2005 study may have been higher because it included cases with multiple deceptions.

Third, the difference in accuracy between the two studies (64.2% vs 73.0%) may fall within normal error variance, and is therefore not meaningful. Indeed, a test of proportions for the differences in decision accuracy between the two studies did not achieve statistical significance (z = 1.36, ns). This does not mean that there is no true difference, only that we did not detect any with the sample size of 100 cases.

One of the central purposes of the present study was to look at interrater agreement. We took steps to recruit polygraph examiners with considerable experience conducting the RI screening test, but in doing so our resultant sample of examiners was quite homogenous. We acknowledge that this sample may not be generalizable beyond government. Our examiners had the same RI training, had run the same kind of RI cases in the field with the same RI technique, and had worked within a single highly standardized system, conditions that may not well represent examiners working elsewhere. Under these favorable conditions, it was anticipated that our findings might surpass the level of interrater agreement found in other polygraph studies.

This was not the case, however. The current data suggest that pairs of experienced examiners can agree on RI cases nearly two times out of three. While significantly greater than chance, such unremarkable performance was not anticipated. No Opinion decisions did not account for most of the disagreements, either. A post hoc analysis showed unanimity among all scorers occurred only 33 times, 32 of them for SR, one for NSR, none for a No Opinion. There was an opposite decision (NSR vs SR) by at least one blind scorer in 51 of the 100 cases. The high rate of opposite calls is not found in the literature for any other polygraph technique. If replication bears a similar result, there could be practical implications for the use of the RI screening test.

We were unable to locate published reports on interrater agreement for RI cases. Nevertheless, our findings of paired agreement of about .60 for the RI screening test were a marked departure from reliability statistics of other polygraph screening techniques. For example, in the APA's Report of the Ad Hoc

Committee on Validated Techniques (2011), screening techniques for which there were reliability data showed an average agreement between .81 and .97 for versions of the Test for Espionage and Sabotage and the Air Force Modified General Question Test (page 244, Table 6). Because validity cannot exceed reliability, the modest reliability of human evaluators with RI data is cause for concern.

It is premature to consider our findings more than tentative, pending independent confirmation by other reliability studies for the RI. We propose, however, that our homogenous sample of blind scorers worked in favor of higher reliability, and that it would be optimistic to believe future studies with field data would result in more positive reliability statistics. Accepting the present findings for a moment, we considered why the RI may produce significantly lower agreement among blind scorers.

A candidate hypothesis is rooted in the method of analysis that comes with this technique, that is, reliance on global impressions of the data. Virtually all other polygraph techniques have numerical scoring systems, approaches that rarely result in diametrically opposed decisions between scorers. By quantifying and summarizing polygraph data in a formal system, scorers tend to enjoy high interrater and intrarater agreement. Global evaluation, in contrast, with no scoring system, invites more subjectivity into the process. With no systematic approach to data analysis, global evaluators can come to rely more on individual experience, skill and biases. Under those conditions, erosion of interrater agreement is not only more likely, but perhaps inevitable as differences in experience, skill and bias increase along with the size of the polygraph screening program.

RI analysis is further complicated by the lack of response benchmarks which other techniques have in the form of probable- and directed-lie comparison questions. Absent these benchmarks, RI data necessarily relies on more ambiguous rules or assessments, leading to more frequent differences in conclusions, lower reliability, and ultimately to lower accuracy.

It seems eminently reasonable that interrater agreement could increase by the development and implementation of a valid scoring system. To date there have been no reports of any effective manual scoring system for the RI screening test. However, there are prototype automated algorithms that offer promise. Two systems were developed under US government contract to provide decision assistance to field examiners (Cook, Kennard & Almond, 2003; Harris & McQuarrie, 2002). The latter system drew cases from the same database as used in this project, and produced accuracy about 9% higher than our human evaluators (z = 1.34, ns). The laboratory work of Kircher, Woltz, Bell and Bernhardt (2006) with the RI screening test offered a different algorithmic approach but accuracy similar to the other algorithms. The chief advantage to automated analysis is that it offers perfect reliability. Human scorers are less consistent, and global evaluation appears to compound the reliability problem. It would be of great interest to determine whether automated decision assistance improves agreement and accuracy for the RI screening test, and we would recommend future study in this area.

Study Limitations

Our evaluators were all experienced practitioners with similar field experiences. The level of agreement achieved in this study may only generalize to experienced RI users accustomed to working within a quality control program, and may be different in other settings.

The RI cases had four relevant questions which were presented in quasi-random order over three charts. Variations from this testing protocol may produce different results.

The population from which the cases were drawn consisted primarily of African American females. Demographic information for individual cases was no longer available for our sample, and therefore statistical treatments of those variables were not possible. Previous research suggests there are group differences in physiological responding that correspond with gender and race (Anderson & McNeilly, 1991). Those differences have not been shown to affect decision accuracy with probable-lie tests, but the effect of demographic variables on the RI screening test is unknown.

The routine destruction of study documentation also denied us the opportunity to conduct other analyses. For example, in the Krapohl et al. (2005) study it was determined that the original polygraph examiner gained many posttest admissions of marijuana use that were missed by urinalysis, and the urine test detected marijuana use the polygraph examiner missed in a small percentage of cases. If the current sample overrepresented those cases in which the original examiner detected the marijuana use at the expense of those in which he missed it, our findings may have produced a more optimistic estimate of decision accuracy. With no records available as to the basis for the ground truth, it remains an open question.

Summary

This study showed the RI screening technique to meet the validity requirements of the APA Standards of Practice, consistent with the findings of Krapohl, Senter and Stern study (2005). Added here were metrics for interrater agreement. Average agreement was significantly greater than chance, but substantially lower than that reported for other screening techniques. Not all pairs of scorers showed above-chance agreement. The RI screening test as described here and its many local variations are used in applicant testing by a large number of examiners, particularly in government and law enforcement. We recommend other studies to assess the RI screening test, as well as the development of analytical approaches that might improve validity and reliability.

References

American Polygraph Association (2011). Report of the Ad Hoc Committee on validated techniques. *Polygraph*, 40(4), 196-305.

Anderson, N.B., & McNeilly, M. (1991). Age, gender, and ethnicity as variables in psychophysiological assessment: Sociodemographics in context. *Psychological Assessment*, 3(3), 376-384.

Brownlie, C., Johnson, G.J., & Knill, B. (unpublished). Validation Study of the Relevant/Irrelevant Screening Format.

Cook, D.R., Kennard, R.F., & Almond, J.P. (2003). Pilot Study: Polygraph Decision Support System Using Event Resolution Imaging for the Relevant/Irrelevant Format. Thoughtform Corporation: Bountiful, UT

Harris, J.C. & McQuarrie, A.D. (2002). The Relevant/Irrelevant Algorithm Description and Validation Results. The Johns Hopkins University, Applied Physics Laboratory. DoDPI02-R-0008.

Kircher, J.C., Woltz, D., Bell, B., & Bernhardt, P.C. (2006). Effects of audiovisual presentations of test questions during Relevant/Irrelevant polygraph examinations and new measures. *Polygraph*, 35(1), 25-54.

Krapohl, D., Senter, S. & Stern, B. (2005). An exploration of methods for the analysis of multiple-issue relevant/irrelevant screening data. *Polygraph*, 34(1), 47-61.