EUROPEAN
POLYGRAPH

Volume 9 • 2015 • Number 4 (34)

DOI: 10.1515/ep-2015-0007

Donald J. Krapohl
Walt Goodson
American Polygraph Association
United States

Decision Accuracy for the Relevant-Irrelevant Screening Test: Influence of an Algorithm on Human Decision-Making[1]

Key words: Relevant-Irrelevant Screening Test, Relevant-Irrelevant Test, Accuracy of R-I list, Screening, decision-making

The Relevant-Irrelevant (RI) has been used as a polygraph screening technique by several decades longer than any other. It has demonstrated practical value in prompt-

[1] This is the third in a series of research articles on the RI screening test. The authors would like to express appreciation to the following examiners for participating in this study: Gus Trevino, Derick Walker, Timothy Upright, Daniel Scheel, Jerry Lesikar, Bill Horton, Steven Davis, Eddie Hutchinson, Dana Wickland, Leo Perez, and Eduardo Garza. We also are grateful to Mark Handler for his helpful suggestions to an earlier version of this paper. The first author is a Past President of the American Polygraph Association, and author of the Elsevier textbook *Fundamentals of Polygraph Practice*. The second author is President of the American Polygraph Association, and contributing author to past articles in this publication. The views expressed are the authors' own, and do not necessarily represent those of the Department of Defense, US Government, or the Texas Department of Public Safety. Comments can be sent to Krapohld@gmail.com.

ing self-report from applicants and employees of behaviours of interest to employers. The RI has certain strengths that have made it an attractive alternative (Krapohl & Shaw 2015). With no comparison questions, the RI is not subject to criticisms that the examiner must manipulate the examinee in some fashion to make the technique effective, as probable-lie comparison question techniques may. It is more flexible than most other methods, accommodating from two to five relevant questions in a single test series. The RI may also be more resistant to countermeasures, at least of the type in which examinees induce reactions to comparison questions.

These advantages have kept the RI in widespread practice well after the development and validation of other polygraph screening techniques. However, the existence of these other techniques has brought about comparisons that have not always favoured the RI. For example, the RI does not seem to be amenable to any effective form of manual scoring, a standard for every other screening technique. As such, interpretation of RI data is far more susceptible to individual differences among evaluators. Lenient evaluators can make conclusions highly different from those of conservative evaluators because decisions are based on global impressions of the data, impressions that resist quantification. This was recently demonstrated in a compelling manner in a study where pairs of four experienced polygraph examiners produced opposite opinions in 51 of 100 field RI cases (Krapohl & Rosales 2014).

Another challenge to the RI has been in the finding that its decision accuracy does not compete well against other techniques: Other validated screening techniques have shown higher decision accuracy. This should not be entirely surprising given that a test's reliability sets the limit for its validity, and the RI's unimpressive reliability *should* naturally translate to a lower accuracy. To determine whether RI could possibly achieve accuracy comparable to other screening techniques, it seems reasonable as a first step to try and resolve RI's reliability problem. This would require some form of data quantification.

Past efforts to develop a manual scoring system for RI charts have not produced satisfactory results (Ansley & Weir 1976, Krapohl, Senter & Stern 2005). With computer polygraphs now standard across the profession, an obvious alternative to manual scoring could be the use of an automated algorithm. How automated analysis might improve reliability is self-evident: software analyses the data the same way each time, and is unaffected by human foibles such as bias, mood, attitudes, fatigue, limited experience, and/or poor training. Algorithms are far more reliable than humans in managing complex data needed to form decisions. It seems reasonable, therefore, that an algorithmic approach could increase reliability of data interpretation, and set the stage for potentially greater decision accuracy. a prototype of an RI algorithm

was developed in the early 2000s (Harris & McQuarrie ca 2001). Its unpublished preliminary results were promising, but the algorithm was never subject of an independent assessment, nor was it released to the field. Its influence on decision-making was, therefore, untested.

Algorithms are not without shortcomings. What polygraph algorithms may do less well is identification of artefacts and assessment of stability in the data, an area where humans may have an advantage. Humans are fairly proficient at pattern recognition, a skill helpful in establishing context for the assessment of individual responses, and they can adjust their decision process accordingly. Humans may also be better at detecting idiosyncratic tell-tale patterns for individual examinees in a way that allows a more informed decision than would arise from a strictly statistical one based on group averages.

Based on the unique and disparate capabilities of humans and machines, it is our hypothesis that the combination of human evaluation and algorithmic analysis may improve decision accuracy and interrater reliability for the RI over either human or machine alone. There is no published evidence to suggest whether this is a valid hypothesis, as the polygraph literature is totally silent on human-machine decision making. Our expectation of better accuracy using human-machine collaboration is based on findings from a totally different field, that of weather prediction.

Weather forecasting is heavily relied on in a range of endeavours such as agriculture, travel, construction, civil defence, and others. Consequently, a great deal of attention and funding, has been directed toward developing and assessing weather models, and the place for human decision-making in the overall process. For example, a study of the combination of statistical decision-making with human refinement in weather prediction was undertaken by Carter and Polger (1986). In their 20-year assessment of historical weather data, Carter and Polger found that the most accurate prediction of weather events came when forecasters modified the statistical prediction with factors they knew were influenced by their locality. The National Weather Service models provided the larger framework, and meteorologists' personal experience and knowledge modified those predictions to produce the highest accuracy for their regions of the country. In his book *The Signal and the Noise* (2012), Nate Silver noted that the contribution of human input to computer weather forecasts has been relatively flat at about 25% for precipitation and 10% for temperature, even as forecasting and computer modelling has improved substantially over the decades. The weather modelling literature makes a strong case in that domain that human adjustments of statistical predictions are better than either method alone.

Polygraph decision-making, like weather prediction, is a combination of statistical averages and local variations. In other words, polygraph examiners apply standardised scoring methods to their data, but may consider factors concerning each individual that may adjust their decisions. As an example, if the polygraph data indicate deception, but an examiner observes that his examinee has become emotionally distraught during testing, a statistical decision of deception may be suspended. Examiners may also informally or unconsciously consider extra-polygraphic information, adjusting their scores in the direction they judge most likely correct (Elaad, Ginton, & Ben-Shakhar 1994), which can move decisions to or from inconclusive when the physiological data are ambiguous. Barland (1988) also suggests that an examiner's knowledge and appreciation of base rates may make them more cautious in ascribing deceptive intent when base rates are low.

Though the polygraph literature shows some research that compares human decision accuracy against that of an algorithm (Blackwell 1999, Krapohl & McManus 1999) we did not find any on the influence of algorithm outputs on the validity and reliability of human polygraph decisions. Moreover, the literature is completely silent on the melding together of human and statistical assessments to form polygraph decisions. Our two research questions were therefore straightforward: what is the influence of algorithmic results on human decision-making, and; can validity and reliability be improved by integrating human and algorithmic results to form a single polygraph result? We took advantage of access to skilled polygraph examiners in a large state agency, a substantial archive of verified field RI screening cases, and the results of a prototype RI algorithm to test these questions.

Method

Cases

In the late 1990s, the US government sponsored a research study on the RI screening test using job applicants for security positions at a large metropolitan airport. Confirmation of ground truth was established by record checks and urinalysis. There were four relevant topics covered in those cases: convictions or fines for traffic violations in the State of Georgia in the previous seven years, having been granted bankruptcy in the previous seven years in the State of Georgia, having used marijuana in the previous 30 days, and having been convicted of a felony in the State of Georgia. a description of the original RI research study that produced this archive of cases can be found in Krapohl, Senter, and Stern (2005), and is not recounted here.

We randomly chose electronic polygraph charts of 50 confirmed deceptive cases in which the examinee had been deceptive solely to the question regarding marijuana use, and another 50 from cases in which the examinees had been truthful to all of the relevant questions. The rationale for this sampling approach was fully explained in Krapohl and Rosales (2014), but in brief, it was based on minimum criteria for the quantity and fidelity of the data. The use of cases where there was deception only to a single question was based on earlier findings from Krapohl, Senter, and Stern (2005) that decision accuracy was higher when there were multiple deceptions in a single case, but that multiple-deceptions represented a small minority of all cases. We concluded that the use of cases with multiple deceptions in the sample would limit generalisability of the results, and we eliminated those cases. There was no attempt to exclude single-deception cases that had been selected in previous studies.

All cases had three charts and at least four presentations of each of the four relevant questions. In the original file some cases had four charts, the last being called a "clearing chart." Clearing charts are used in cases where the testing examiner determined that there were indications of deception on a question in the first three charts, and the examiner used the fourth chart solely to "clear" the remaining questions. Because the presence or absence of a clearing chart would indicate the opinion of the original examiner, those charts were not made available to the blind scorers in this study. As a result, all scorers saw only the first three charts of each RI case.

The physiological data were converted to PDFs, with one file per case, and randomly numbered from 1 to 100 for the first phase of the study, and re-randomised and numbered from 101 to 200 for the second phase.

Blind Evaluators

There were 11 volunteer evaluators from the Texas Department of Public Safety. All held polygraph licenses in the State of Texas. Their average field experience was 6.1 years (ranging from 1 to 20), and their average annual polygraph-related training was 82 hours. All had received training in RI screening techniques and global test data analysis. Four of the examiners had experience conducting RI cases in the field. The examiners all averaged 12 screening examinations in the field per month and all of these examinations were vetted through a 100 percent quality assurance programme.

Algorithm

Johns Hopkins University Applied Physics Laboratory developed the prototype algorithm using the cases collected in the RI project (Harris & McQuarrie ca 2001.) The features used in the algorithm relied on derivatives and weighting of reordered data

points for which it would be difficult or impossible to map to traditional physiological features (e.g., the absolute derivative of reordered abdominal respiration data between the Xth and XXth percentile in time window X to XX seconds and a proportionate weighting of X%). The details of the algorithm features are beyond the scope of this project, and the algorithm itself is proprietary, so the features will not be described further. Decision accuracy of that algorithm in a previous study indicated a mean of 73% for cases that included single and multiple deceptions (Krapohl, Senter & Stern 2005).

The prototype RI algorithm was run against all the 100 cases, and the results were recorded in a spreadsheet. Artefact management was handled by the algorithm, with the first author making corrections only for erroneous question labels. The algorithm results for each case were provided to the blind evaluators only in the second phase of the study. a decision was reached to consider probabilities of deception rendered by the algorithm as lower than 0.30 as No Significant Responses (NSR), while those above 0.70 were called Significant Responses (SR). All the others were called No Opinion (NO).

Data Analysis

Some blind evaluators reported decisions of Deception Indicated (DI), No Deception Indicated (NDI), and Inconclusive, while others used the equivalent labels used in screening, which are Significant Reactions (SR), No Significant Reactions (NSR) and No Opinion (NO). To permit easier comparison to earlier RI studies with these cases, the decisions are reported here as SR, NSR, and NO, respectively, to be consistent with the labelling convention used in Krapohl and Rosales (2014), and Krapohl, Senter, and Stern (2005).

A spreadsheet was constructed in Microsoft Excel®, and the decisions were recorded. Decisions were coded as -1 for SR, +1 for NSR, and 0 for NO.

Procedure

The study was conducted in two phases. In Phase 1, examiners were told to evaluate individually the RI charts using global analysis in the manner they were accustomed, and to record their decisions[2] on a data sheet. The examiners were instructed to make decisions by case, not by individual test question. Examiners were not informed of

[2] The RI is typically the first in a multi-step screening process, wherein all reactions to RI test questions are resolved by further interviewing and more focused testing. Decisions of SR or NO are not appropriate to RI screening alone, but the labels have been used here for convenience.

ground truth or of the test questions. They were asked to work independently, and – to avoid sharing their opinions about the cases with other evaluators – they were also given 30 days to complete the task, with extensions if requested. Examiner instructions are found in Appendix A.

Three months after the receipt of all results from the blind evaluators, the procedure was repeated as Phase 2. In that iteration the file numbers were re-randomised and re-labelled, and the algorithm results were posted on the first page of the PDFs with the physiological data. Examiners received the same instructions as in the first phase, with an extra paragraph added in the instructions that read:

Please review the PolyScore results on the first page of the PDF of each case. Previous research has found this algorithm performs as well as a competent examiner in blind evaluation of RI charts. You are not obligated to use the algorithm results, but only to consider them in formulating your own opinion.

Results

Phase 1: Global Analysis

Table 1 lists the average accuracies for the 11 blind evaluators of the 100 RI cases using global analysis. Average decision accuracy was 74.4% for deceptive cases, and 52.7% for truthful cases, for an overall average of 63.6%. Decision accuracy for deceptive cases was significantly greater than chance ($z = 2.51$, $p < 0.05$), but not for truthful cases ($z = 0.27$, ns). When No Opinion decisions were excluded, overall decision accuracy was 66.9%, which was also significant greater than chance ($z = 2.86$, $p < 0.05$). Collectively, the 11 examiners had a higher decision accuracy for deceptive cases than for truthful cases ($z = 2.25$, $p < 0.05$).

Table 1. Average percentages of correct, incorrect, and No Opinion (NO) decisions by ground truth, with and without NO decisions, for 11 blind evaluators of 50 deceptive and 50 truthful RI cases.

	Deceptive	Truthful	Overall	w/o NO
Correct	74.4	52.7	63.6	66.9
Incorrect	23.5	39.5	31.5	33.1
Inconclusive	2.2	7.8	5.0	

Table 2 shows the proportion of agreement between each evaluator's decisions and ground truth, and between each pair of evaluators. Interrater agreement ranged from 42.0% to 80.0%, with a mean of 60.8%. Only 4 of 55 paired agreements did not exceed chance expectancy. There was unanimous agreement among all evaluators in 11 of the 100 cases, and of those, 10 were decisions of deceptiveness. Only one of those unanimous decisions was in error, a false positive. Of the 50 deceptive cases there were 38 in which at least one evaluator made a decision that opposed (NSR against SR, or SR against NSR) that of the remaining 10 evaluators. In the 50 truthful cases, 48 had at least one opposite decision among the evaluators.

In Phase 1, the evaluators did not have access to the RI algorithm. However, the algorithm's decisions were compared to their Phase 1 decisions to provide a benchmark for the degree of influence the algorithm would have when the evaluators were exposed to its decisions in Phase 2. Overall average agreement between the individual evaluator decisions and the algorithm decisions was 59.9%. For deceptive cases the average was 64.8%, and 54.0% for truthful cases.

Table 2. Percentage of agreement for decisions on 100 RI cases between ground truth and each examiner, and between pairs of examiners. All percentages were greater than chance ($p<0.05$) except those marked *. Chance agreement = 33.3%.

	Examiner										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Ground Truth	73	64	64	58	62	59	55	60	62	81	61
Examiner 1		60	68	62	68	60	64	62	63	80	64
Examiner 2			68	43*	68	67	67	59	61	62	70
Examiner 3				49	73	68	71	46*	65	61	69
Examiner 4					44*	38*	42*	62	57	64	50
Examiner 5						68	64	56	63	70	58
Examiner 6							60	51	62	55	65
Examiner 7								52	58	62	62
Examiner 8									57	64	49
Examiner 9										67	63
Examiner 10											63

For the entire 100-case sample the average SR rate was 56.9%, 38.1% NSR, and 5.0% NO. The rate at which the 11 evaluators made SR, NSR, and No Opinion decisions varied substantially (See: Table 3). With a base rate of 50 deceptive cases for the 100-case sample, evaluators made from 25 to 78 decisions of SR. With the same base rate of truthful cases evaluators made from 22 to 67 NSR decisions. Evaluators made No Opinion decisions in 0% to 13% of the 100 cases.

Table 3. Number of SR, NSR, and No Opinion decisions of the 11 evaluators of 100 RI cases.

	Examiner										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
SR Decision	52	78	66	25	72	71	61	47	52	46	56
NSR Decision	47	22	34	67	26	23	32	45	35	53	35
NO Decision	1	0	0	8	2	6	7	8	13	1	9

Discussion

Decision accuracy in Phase 1 of this study (63.6%) corresponded well with the findings of Krapohl and Rosales (2014) of 59.3%, and a reanalysis of Krapohl, Senter, and Stern (2006) of single-deception RI cases showing 63.0% decision accuracy. Similarly, the average proportion of paired agreement among evaluators in Phase 1 of 60.8% was highly similar to that of Krapohl and Rosales (2014) at 59.7%. These comparable findings across different data sets using different scorers point to lacklustre performance of global evaluation when it is used alone in RI cases.

Phase 2: Global Analysis + Algorithm

Table 4 lists the average accuracies for the blind evaluators when they had access to algorithm results when conducting their global assessment of the 100 RI cases. Average decision accuracy was 72.5% for deceptive cases, and 60.7% for truthful cases, for an overall average of 66.6% that was significantly greater than chance ($z = 2.38$, $p < 0.05$). As was found in Phase 1, decision accuracy for deceptive cases was significant ($z = 3.27$, $p < 0.05$), but it was not so for truthful cases ($z = 1.52$, ns). When No Opinion decisions were excluded, overall decision accuracy was 75.3%, which was also significantly greater than chance ($z = 4.98$, $p < 0.05$). There was no difference in accuracy between truthful and deceptive cases ($z = 1.77$, ns).

Table 4. Average percentages of correct, incorrect, and No Opinion (NO) decisions by ground truth, with and without NO decisions, for 11 blind evaluators with access to algorithm results of 50 deceptive and 50 truthful RI cases.

	Deceptive	Truthful	Overall	w/o NO
Correct	72.5	60.7	66.6	75.3
Incorrect	18.4	25.2	21.8	24.7
Inconclusive	9.1	14.0	11.5	

By way of comparison, for deceptive cases the algorithm results were 68% correct, incorrect for 16%, and NO for 16%. For truthful cases algorithm results were correct for 72%, incorrect for 14%, and NO in 14% cases. Detection of deceptive ($z = 2.91$, $p < 0.05$) and truthful cases ($z = 3.19$, $p < 0.05$) were above chance, and there was no difference in detection rates between the two types of cases ($z = 0.62$, ns). Overall correct decisions were 70% with NOs included, and 82.3% without NOs.

The percentages of agreement between each evaluator's decisions and ground truth, and between each pair of evaluators is listed in Table 5. Interrater agreement ranged from 44.0% to 88.0%, with a mean of 64.6%. All but one paired-agreements did not exceed chance expectancy. There was unanimous agreement among all evaluators in 18 of the 100 cases, and of those, 15 were decisions of deceptiveness, three for truthfulness. All unanimous decisions were correct. Of the 50 deceptive cases there were 29 in which at least one evaluator made a decision opposite (NSR vs. SR, either way) to the remaining 10 evaluators. For the 50 truthful cases, 39 had at least one opposite decision among the evaluators.

Overall average agreement between individual evaluator decisions and the algorithm decisions was 66.6% ($z = 4.71$, $p < 0.05$). The average was 72.5% for deceptive cases ($z = 4.02$, $p < 0.05$), and 60.7% for truthful cases ($z = 2.74$, $p < 0.05$).

Table 5. Percentage of agreement for decisions on 100 RI cases between ground truth and each examiner, and between pairs of examiners. All percentages were greater than chance ($p < 0.05$) except that marked *. Chance agreement = 33.3%.

	Examiner										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Ground Truth	71	68	72	63	66	61	64	64	68	73	63
Examiner 1		67	67	68	55	50	63	66	70	64	63
Examiner 2			66	74	64	56	70	75	80	61	88
Examiner 3				62	67	66	64	60	62	72	67
Examiner 4					52	44*	65	65	69	62	71
Examiner 5						67	56	54	66	69	61
Examiner 6							57	54	56	66	58
Examiner 7								66	67	66	69
Examiner 8									71	59	72
Examiner 9										61	79
Examiner 10											63

In Phase 2, the average SR rate was 48.9%, 39.5% NSR, and 11.5% NO. The rate at which the 11 evaluators made SR, NSR, and No Opinion decisions is listed in Table 6. Among the 50 deceptive cases, evaluators made from 30 to 70 decisions of SR, and from 24 to 63 NSR decisions. Evaluators made No Opinion decisions in 0% to 20% of the 100 cases.

Table 6. Number of SR, NSR, and No Opinion decisions for 100 RI cases by 11 evaluators who viewed algorithm results.

	Examiner										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
SR Decision	37	45	54	30	70	68	43	43	48	56	44
NSR Decision	63	40	44	48	24	18	37	47	37	38	39
NO Decision	0	15	2	22	6	14	20	10	15	6	17

Discussion

A comparison of Phase 1 and Phase 2 data hints of increases in decision accuracy (63.6% vs 66.6%, respectively) and interrater agreement (60.8% vs 64.6%, respectively), but the differences proved to be statistically insignificant. Given that the al-

gorithm accuracy was higher than examiner decisions that considered the algorithm results suggested the use of automated analysis as decision support may not be its best role in the decision process.

Post Hoc: Algorithm + Examiner

By itself, the algorithm had a decision accuracy higher than the average accuracy of humans even when they had access to the algorithm decisions. We then considered whether a two-step process, wherein the algorithm would be afforded the first assessment of the data would offer a better approach, and only in cases where the algorithm was unable to decide, the decision would be submitted to evaluation by the examiners. Fortunately, this hypothesis could be successfully tested with the existing data.

We substituted algorithm decisions for those of the examiners in Phase 1, where the examiner had not seen the algorithm results. In cases where the algorithm had produced inconclusive results, the examiners' global evaluations were retained. While this approach to decision making would be expected to improve interrater agreement, inasmuch as all but the 15 inconclusive algorithm decisions would be identical, whether decision accuracy would show a benefit was an open question.

Results

The Algorithm + Examiner (A+E) decision rules boosted overall accuracy, the improvement coming almost exclusively from the correct identification of truthful cases. a reduction of overall error rates fell short of significance. The effect of the A+E decision rules had no effect on deceptive cases. Pair agreement increased, and there was a significant reduction in opposite results arising from among pairs of scorers. See Table 7.

As Table 7 shows, decision accuracy for the A+E method was significantly higher than the results by examiners alone. Virtually all of the benefit came from the higher accuracy with truthful cases. There were significant improvements in reliability and opposite decisions as well.

Table 7. Decision accuracy, interrater reliability, and statistical probabilities of the differences for two decision processes for 11 scorers of 100 RI screening cases.

		Original Examiner Decisions (%)	Decisions of A+E (%)	Diff p
Overall				
	Correct	63.5	79.5	<0.01
	Error	32.5	19.5	0.051
	No Opinion	5.0	1.0	0.10
Deceptive Cases				
	Correct	74.4	79.8	0.36
	Error	23.5	19.8	0.52
	No Opinion	2.2	0.4	0.26
Truthful Cases				
	Correct	52.7	79.1	<0.001
	Error	39.5	19.3	<0.01
	No Opinion	7.8	1.6	<0.05
Reliability				
	Paired Agreement	60.4	93.0	<0.001
	Opposite Decisions	29.8	6.4	<0.001

To determine whether these findings would generalise to a new sample, we reanalysed the decisions of four experienced examiners for 100 RI screening cases in a previous RI study (Krapohl and Rosales, 2012) using the steps described earlier. Table 8 compares the original decisions with those of the new method. There was an increase in overall correct decisions with the A+E method, but it did not achieve statistical significance. Again, most of the improvement appears to be attributable to the better discrimination of truthful cases. There was, however, a significant difference for errors with deceptive cases, where the original examiner's errors were lower than those of the A+E method.

Table 8. Decision accuracy, interrater reliability, and statistical significance of the differences for two decision processes using examiner decisions from Krapohl and Rosales (2012).

		Original Examiner Decisions (%)	Decisions of A+E (%)	Diff p
Overall				
	Correct	59.3	72.0	0.06
	Error	32.5	26.8	0.38
	No Opinion	8.3	1.3	<0.02
Deceptive Cases				
	Correct	81.5	74.0	0.20
	Error	12.0	25.5	<0.01
	No Opinion	6.5	1.3	0.06
Truthful Cases				
	Correct	37.0	70.0	<0.001
	Error	53.0	28.0	<0.001
	No Opinion	10.0	2.0	<0.02
Reliability				
	Paired Agreement	60.5	94.3	<0.001
	Opposite Decisions	24.7	3.5	<0.001

General Discussion

The findings from the reanalysis of the Krapohl and Rosales (2012) data from four examiners are largely in accord with those from the first sample from 11 examiners (Table 7). Both data sets show marked benefits of the A+E method for interrater agreement and a reduction in opposite decisions. They also share a common finding that the benefit is loaded on the detection of truthful cases.

Where they most noticeably differ is in the error rate for deceptive cases. As far as the decisions of the examiners in Phase 1 of this study (Table 7) are concerned, there was a non-significant reduction in false negative errors for those cases when the A+E rules were imposed, whereas there was a statistically significant increase when the same A+E rules were used with the Krapohl and Rosales (2012) examiners (Table 8).

Because screening tests, as the first step in the successive hurdles approach, must be sensitive to deceptiveness this difference warrants attention.

A possible explanation lies in differences among the groups of examiners used in the present study and those in the Krapohl and Rosales (2012) research. The present group of 11 examiners came from a state agency in which the RI screening test is taught, but not often used. In contrast, the four Krapohl and Rosales (2012) examiners were federal examiners with substantial field experience with the RI screening test, and had at one time all worked for the same federal agency using this method. The difference in experience could be offered as a possible source of the disparate performance with deceptive cases.

The evidence of such an experience difference between the sample groups is not straightforward, however. When Tables 7 and 8 are compared they show that both groups had highly similar overall decision accuracy, proportions of paired agreement, and rates of opposite decisions. It would be difficult to discern which group was the more experienced based solely on their overall performance data.

Another explanation may be that one group had a bias in decision-making that would be manifested in the types of errors the group made. For example, if one group of examiners had a bias away from certain kinds of decisions, and the A+E method did not, going from one method to the other could shift the kinds of errors made. In reassessing the Krapohl and Rosales (2012) data, it appears their lower rate of errors with deceptive cases may be attributable in part to a general reluctance of those examiners to render NSR decisions. Clearly, if examiners avoid NSR decisions, false positive errors should occur less often. a comparison of average NSR rates from the Krapohl and Rosales (2012) group to those of the examiners in Phase 1 of the present study found significantly fewer NSR decisions ($z = 2.14$, $p < 0.05$) among the former group, as well as fewer NSR decisions than reached by the algorithm ($z = 4.20$, $p < 0.01$). Three out of the four examiners in the Krapohl and Rosales (2012) group had decision accuracy statistically *lower than chance* with truthful cases. With the algorithm having a dominant effect on decisions in the A+E arrangement, and lacking the bias, it might be anticipated that the Krapohl and Rosales (2012) examiners would have fewer false positive errors without the A+E rules than with them. This may account for the higher false positive error rate in the A+E arrangement versus the decisions of the examiners in the Krapohl and Rosales (2012) study, an effect not observed among the decisions of the 11 examiners in Phase 1. These findings merit future research.

Summary

We found that allowing examiners to make decisions on the RI screening data only when the algorithm produced inconclusive results produced marked improvements in interrater reliability and detection of truthfulness, and a reduction in opposite decisions among examiners. In one comparison we found the A+E method increased overall decision accuracy. In the cross validation sample the improvement missed significance. The types of errors introduced by the A+E method may be different from those examiners make on their own, though overall error rates are similar. Because the algorithm used in this study is not generally available, our accuracy findings in the A+E arrangement are not expected to represent those of the RI screening test as currently practiced in the field. The converging evidence for decision accuracy for the RI screening test using only global analysis is about 62%, with a similar proportion of agreement among independent examiners. Figure 1 summarises the reliability and accuracy findings of the decision data collected in this study.

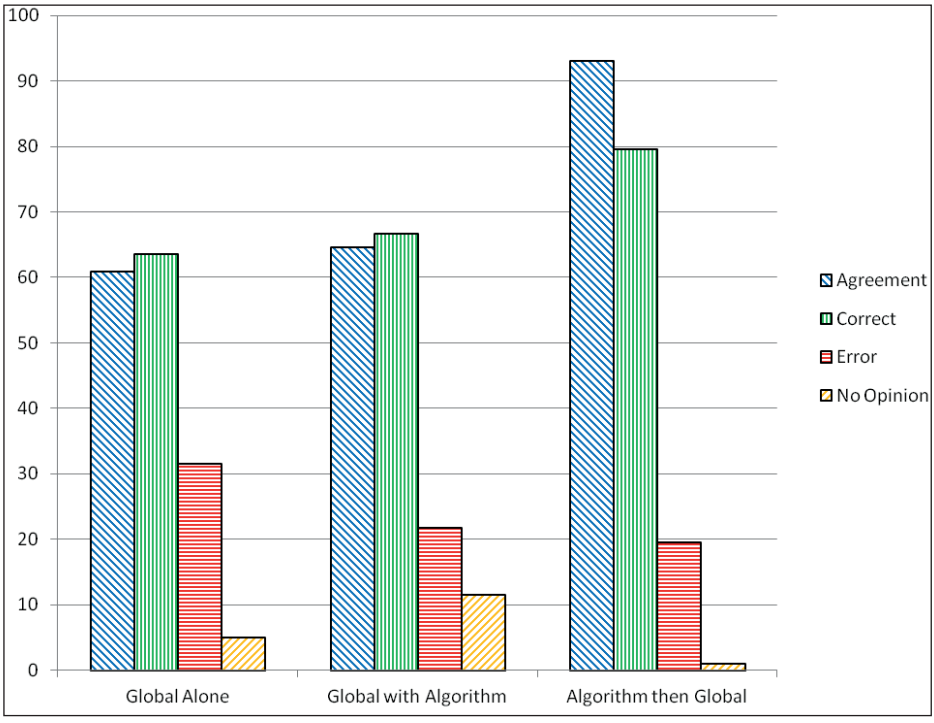


Fig. 1. Summary graph of the three approaches in this study for analysis of RI polygraph charts.

References

Ansley N., Weir R. (1976): a numerical scoring system for Relevant-Irrelevant polygraph tests. Paper presented at the 1976 Annual Seminar of the American Polygraph Association.

Barland G.H. (1988): The polygraph test in the USA and elsewhere. In A. Gale (Ed.) The polygraph test: Lies, truth and science. Sage Publications, London.

Blackwell N.J. (1999): Polyscore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examinations from actual criminal investigations. Polygraph 28 (2), 149–175.

Carter G., Polger P. (1986): a 20-year summary of National Weather Service verification results for temperature and precipitation. Technical Memorandum NWS FCST 31. Washington DC, National Oceanic and Atmospheric Administration.

Elaad E., Ginton A., Ben-Shakhar G. (1994): The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. Journal of Behavioral Decision Making, 7 (4), 279–292.

Harris J.C., McQuarrie A.D. (ca 2001): The Relevant/Irrelevant Algorithm Description and Validation Results. The Johns Hopkins University Applied Physics Laboratory.

Krapohl D., McManus B. (1999): An objective method for manually scoring polygraph data. Polygraph, 28 (3), 209–222.

Krapohl D., Rosales T. (2014): Decision accuracy for the Relevant-Irrelevant Screening Test: a partial replication. Polygraph, 41 (1), 20–29.

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.

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

Silver N. (2012): The Signal and the Noise: Why so Many Predictions Fail – but Some Don't. Penguin Books, New York.

Appendix A

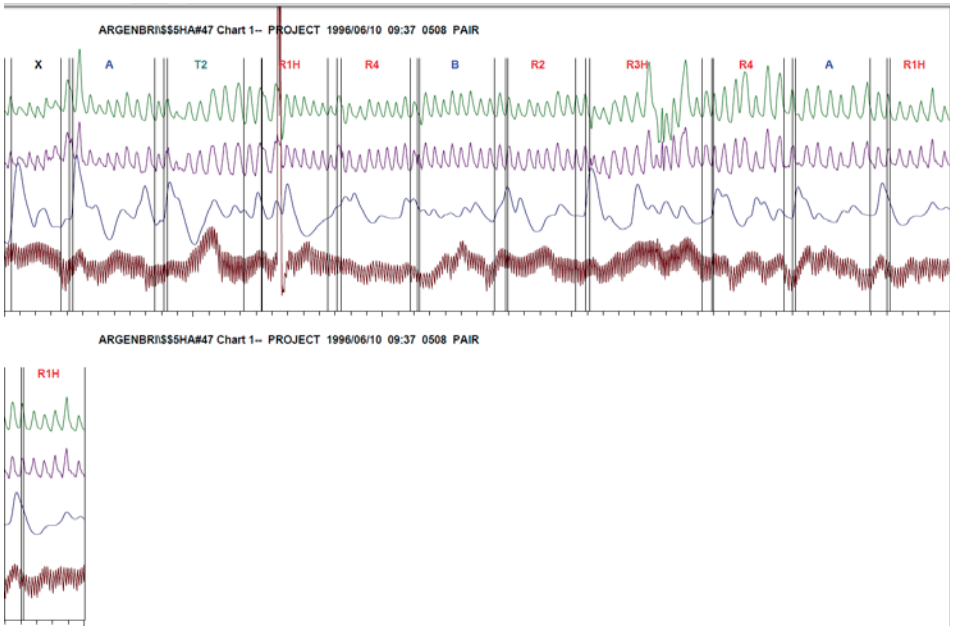
Examiner Instructions

Thank you for volunteering to participate in this project. The data provided by you and your peers will help us in the development of best practices in polygraph screening in general, and the Relevant-Irrelevant (RI) Screening Test in particular. All phases of the project will be completed six months, and we hope to have initial analysis ready for dissemination by summertime. When the analysis is complete, we will issue a formal report of the findings.

Background
This study has two parts. In the first part you are asked to analyze 100 sets of RI charts collected in a field study in the 1990s. Ground truth was established for each of the relevant issues using record checks, urinalysis, and examinee statements. Some, none, or all of the cases come from deceptive examinees. Your task will be to make a decision for each of the 100 cases of Significant Responses (SR), No Significant Responses (NSR) or No Opinion (NO.)

In about three months we will ask you to look at another sample of RI cases in which we will change the instructions of what you need to do. You will make SR, NSR, and NO decisions on those cases, too. When this phase is completed we will compare the accuracy and reliability of your decisions between the two different parts of the study.

Details
Each of the RI cases is in an individual PDF file, one chart per page. Sometimes the charts were too long to print in a single line, and so they may continue in a different line on the same page. Some portions are repeated between the first and second line. See below as an example.



All cases have three charts, each on a different page. The PDF format will allow you to enlarge or reduce the size of the tracings to your preference.

All relevant questions are marked first with an R, then with a number. In most cases they will be R1, R2, R3 and R4. Sometimes examinees made admissions that required modifications of the relevant questions. The modified relevant questions have the letter H at the end of the label, e.g., R1H, R4H, etc.

Irrelevant questions are denoted by a single letter, e.g., A, B, and C. All other labels indicate technical questions, such as T2, R*C2, and I*C2, which are designed to ensure the examinee is capable of responding. They are not comparison questions, and should not be used in that manner.

In virtually every case there are four presentations of each relevant question. Some have only three, and others may have as many as five.

On the Decision Sheet are the numbers 1-100, each of the numbers representing one of the RI cases in the study. The case numbers are in the PDF file label. You should ignore the case notation in the charts themselves. Simply write your decision for the 100 cases as SR, NSR or NO.

Your Rights
This study relies on volunteers. You do not have to volunteer, and if you do volunteer, you have no obligation to finish the study.

You also are ensured confidentiality. Volunteers will be assigned a number that they will use on the score sheets that only you and I (Don Krapohl) will know. When the study is completed, I will send you, and only you, your individual results. There will be no reports that show which data came from which volunteer. Data will anonymised or aggregated in the report submitted for publication.

The benefit to you for participating in this study is that you will receive feedback on your accuracy and reliability in the evaluation of RI screening cases. You will also know that your data were important in the development of best practices in polygraph screening.

Your Responsibilities
Our research project requires each examiner to work independently of the other examiners. This means we ask that you not share any information with other volunteers for six months, or when the study is completed. After that time you will be free to discuss whatever you wish without risking the study data.

Also, please remember that accuracy is more important than speed. We have designated 30 days for the volunteers to complete the analysis of the data, but if you find you need additional time it will be given to you.

Contact Information
If you have any comments or questions, please call me at XXX, or via email at XXX (deleted for this publication).