Addendum to the 2011 Meta-analytic Survey – the Utah Four-Question Test ("Raskin Technique") / ESS

Editorial Staff

APA (2011) published the report of meta-analytic survey of validated polygraph techniques in preparation for evolving standards of practice which require the use of validated techniques in field polygraph practice. Two important aspects of the design of that project were the specification of requirements for inclusion in the study and a definition of what is meant by term polygraph technique. For the purpose of that survey, a polygraph technique was defined as a defined question sequence together with an analysis method. This definition was premised on an assumption that the effectiveness of a polygraph technique is, in part, a function of the recording data for analysis and interpretation and also analysis of the recorded data through a valid and structured process.

Inclusion in the meta-analytic survey required published and replicated studies showing test sensitivity, specificity, false-positive and false negative rates, in addition to the publication of the means and variance of the sampling distributions. The requirement for publication and replication was premised on an assumption that all research samples are biased – they are an imperfect representation of the population – and the fact that sampling statistics, if randomly selected, will converge towards the unknown population parameters according to the central limit theorem (Kwak & Kim, 2017). Researchers in all areas of social science make use of this theorem to develop tests and measures for amorphous phenomena such as personality traits, intellectual functioning, academic achievement, height, weight ,or any population referenced phenomena of interest.

Although the Utah three-question test was included in the meta-analytic survey, the Utah four-question test – sometimes referred to as the "Raskin technique" – was not included due to an absence of published information specific to this format. The Utah four-question test is mentioned by APA (2011), in footnote #50 on page 248, for its structural similarity to the AFMGQT, with an advisement that information can be generalized for the two named formats.

One important difference between the Utah four-question format and the AFMGQT is that the latter is commonly interpreted using the subtotal-score-rule (SSR) whereas the Utah four-question format is commonly interpreted with the grand-total-rule (GTR) or two-stage-rule (TSR). [See Nelson (2018) for a survey of polygraph decision rules.] Underlying the selection of a polygraph decision rule is an assumption as to whether the relevant questions are independent or non-independent (dependent).

Independence, in the scientific context, requires that the questions have no shared source of variance through which factors that influence responses to any question could also influence responses to other questions. The AFMGQT is commonly used in polygraph screening contexts in which relevant questions are formulated to investigate an array of behavioral concerns in the absence of any known incident or allegation, and are commonly interpreted with an assumption of independent criterion variance (notwithstanding that the examinee's attention will remain a potential dependency or influencing factor within a multiple issue exam). For reasons both psychological and statistical (i.e., multiplicity) multiple issue exams cannot provide the same level of accuracy or precision as single issue exams. However, multiple issue exams are useful in polygraph screening programs.

In contrast, the Utah four-question format is used as an event-specific diagnostic polygraph format – used to investigate the verac-

ity of an examinee's statements regarding a known incident. Because all of the relevant questions involve a single known incident or allegation, an assumption of independence is unfounded. A convenient and useful aspect of the Utah four-question single issue test is that any combination of primary relevant and secondary relevant questions (i.e., weak relevant, evidence connecting, guilty knowledge, helping, planning, participating, and role descriptive questions) is permitted. That is, field examiners are free to use any type of relevant question that best suits the circumstances and needs of the investigation. As a matter of practice at least one of the relevant questions is commonly a primary relevant question that describes the examinee's behavioral involvement the topic of the investigation. This is, in many ways, similar to the formulation of questions for the AFMGQT versions 2. The core aspects of the sequence itself [CQ, RQ, RQ, CQ, RQ, RQ, CQ] is structurally identical for the Utah four-question test and the AFMGQT. Another similarity for the two formats is that all CQs and RQs are rotated in a random or pseudo-random manner for each iteration of the question sequence.

Regardless of the combination of primary and secondary relevant questions, the Utah four-question test is interpreted with an assumption of shared criterion variance among the RQs – the examinee was either involved or not involved in the allegation or incident under investigation. An assumption of non-independent criterion variance forms the basis for the use of the GTR or TSR, and simplifies the assumptions and requirements both psychologically and statistically. The examinee is not subject to divided attentional demands, because all relevant questions pertain to the same incident or allegation. Multiplicity effects are not a factor when using the GTR, and are reduced through the use of a statistical correction when using the TSR.

Another important aspect of the Utah four-question format is that the use of four relevant questions, instead of three or two, will mean that the test result will be based on more information compared to three-question and two-question test formats. As a general principle, use of more data leads to increased precision or accuracy of quantitative conclusions. This is related to the *law-of-large-numbers* (Dekking, 2005; Révész, 1968) which holds that the frequency of occurrence of a random event converges towards its probability as the number of trials increases. It is the reason that larger samples are preferred to smaller samples – they can, if randomly selected, more closely approximately an unknown population parameter.

With the Utah four question test, the amount of data for a polygraph test with four RQs and up to five iterations of the question sequence is more than three times that of a test with three iterations of two RQs. The result of this is an increased in both test sensitivity and specificity, with a corresponding reduction in inconclusive results and increase in overall precision Another result is that the Utah four-question test can be more robust, and less vulnerable, in the context of difficult test data.

Raskin, Honts, Nelson and Handler (2015) published the results of a Monte Carlo study on the Utah four-question test, including both ESS and seven-position scores. Seed data for the Monte Carlo were N=100 exams from the University of Utah. Results were statistically undifferentiatable for the two scoring methods. For ESS scores with the TSR using alpha = .05/.05 for deception and truth-telling the unweighted accuracy rate for five iterations of the question sequence was .949 with an unweighted inconclusive rate of .020.

Nelson (2018) published a second study on the Raskin technique using data from th DoDPI confirmed case archive. Examinations were a sample of N=30 confirmed field cases that were conducted using the AFMGQT format. This format was described earlier as structurally similar to the Utah four question test; all cases consisted of three iterations of a question sequence that included a combination of primary and secondary relevant questions. Scored were obtained using an automated version of the ESS-M, and results were classified as deceptive or truthful using the TSR with alpha=.05/.05 deception and truthtelling. Unweighted accuracy was .929 with an inconclusive rate of .033.

Figure 1 shows a mean and standard deviation plot of the scores of the sampling distributions of the included Utah Four-Question

(Raskin) Technique studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not statistically significant *[F (1,88) = 0.673, (p = .414)]*, nor was the main effect for sampling distribution *[F (1,88) = 0, (p = .993)]*. One-way ANOVAs showed no significant differences in the scores of the two studies for either the deceptive samples *[F (1,44) = 0.018, (p = 0.895)]* or truthful samples *[F (1,44) = 0.02, (p = 0.888)]*.

Figure 1. Mean and standard deviations for the scores from truthful and deceptive samples with the Utah four-question format.

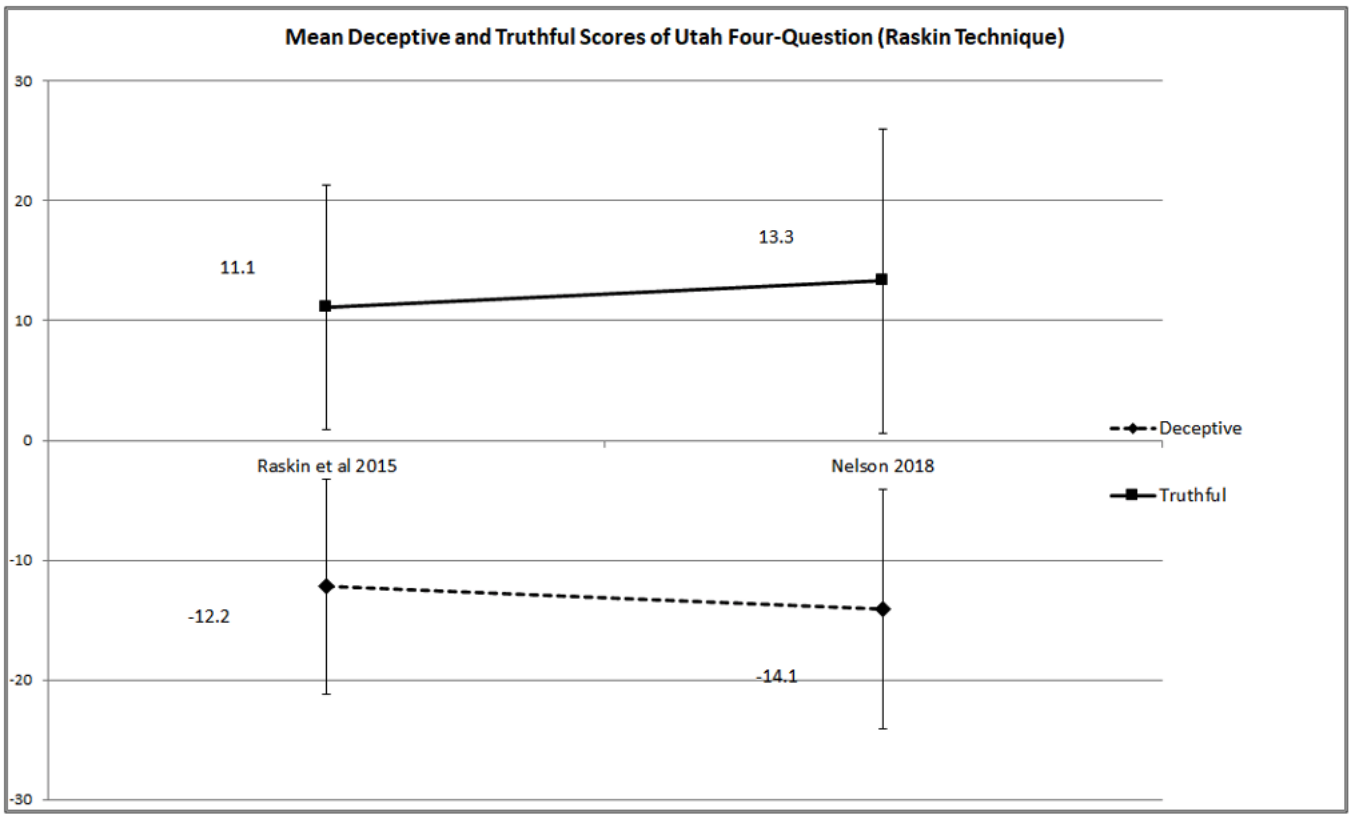

Table 1 shows the summary of the two combined on the Raskin technique.

Table 1. Summary of studies on the Raskin technique.	
Number of usable Studies	2
Total N	130
N Deceptive	65
N Truthful	65
Number of Examiners/Scorers	2
Total Scores	130
D Scores	65
T Scores	65
Mean D	-12.638
StDev D	10.154
Mean T	11.608
StDev T	9.854

Table 2 shows the profile and statistical confidence intervals for the criterion accuracy metrics.

Table 2. Criterion accuracy and confidence intervals for the Raskin technique.	
Unweighted Average Accuracy	.944 (.021) {.897 to .984}
Unweighted Inconclusives	.031 (.026) {.010 to .092}
Sensitivity	.923 (.033) {.852 to .984}
Specificity	.908 (.036) {.831 to .971}
FN Errors	.046 (.026) {.010 to .104}
FP Errors	.062 (.030) {.014 to .125}
D-INC	.031 (.021) {.010 to .078}
T-INC	.031 (.021) {.010 to .078}
PPV	.938 (.031) {.871 to .986}
NPV	.952 (.027) {.892 to .990}
D Correct	.952 (.027) {.893 to .990}
T Correct	.936 (.031) {.871 to .986}
Detection Efficiency Coefficient	.875 (.041) {.788 to .949}

Table 3 shows a summary of the individual studies.

Table 3. Summary of individual studies on the Raskin technique.		
Study	Raskin et al., (2015)	Nelson (2018)
Sample N	100	30
N Deceptive	50	15
N Truthful	50	15
Scorers	1	1
D Scores	50	15
T Scores	50	15
Total Scores	30	30
Mean D	-12.2	-14.1
StDev D	10.2	10.0
Mean T	11.1	13.3
StDev T	9.0	12.7
Unweighted Average Accuracy	.949	.929
Unweighted Inconclusives	.020	.067
Sensitivity	.940	.870
Specificity	.920	.870
FN Errors	.040	.067
FP Errors	.060	.067
D-INC	.020	.067
T-INC	.020	.067
PPV	.940	.929
NPV	.958	.929
D Correct	.959	.929
T Correct	.939	.929

The combined decision accuracy level of the Utah four-question test ("Raskin technique") studies, weighted for sample size and number of scorers, was .944 with a combined inconclusive rate of .031. The detection efficiency coefficient, calculated as the correlation between the categorical result coded as [+1, 0, -1] and the criterion state for each case coded as [+1, -1] was .875 with a 95% confidence interval from .788 to .949.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. [Electronic version] Retrieved September 15, 2020, from http://www .polygraph.org.

Dekking, Michel (2005). *A Modern Introduction to Probability and Statistics.* Springer.

Nelson, R. (2018). Credibility assessment using Bayesian credible intervals: a replication study of criterion accuracy using the ESS-M and event-specific polygraphs with four relevant questions. *Polygraph and Forensic Credibility Assessment*, 47 (1), 85-90.

Kwak, S. G. & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. Korean *Journal of Anesthesiology.* 70(2), 144–156.

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

Raskin, D. C., Honts, C., Nelson, R. & Handler, M. (2015). Monte Carlo estimates of the validity of four-relevant-question polygraph examinations. *Polygraph 44(1)*, 1-27.

Révész, P. (1968). *The Laws of Large Numbers.* Academic Press.