Short Report: A Test of the ESS with Two-Question Field Cases

Donald J. Krapohl[1]

The Empirical Scoring System (ESS; Nelson, Krapohl & Handler, 2008) was developed with the goal of improving inter-scorer reliability while also establishing error estimates for the Zone Comparison Technique at specified cutoff scores. Much of the effort to boost inter-scorer reliability came from simplification: The system reduced the scoring features to the three "Kircher features," employed a three-position scoring regimen, and used only three scoring rules. To establish cutoff scores, statistical analyses of norm data were used to find a balance of sensitivity and specificity which combined for the highest accuracy, and then the cutoff scores were tested on new data sets. The initial report was encouraging, and given the simplicity of the ESS approach, appeared at first impression to be a method more easily taught to polygraph students. In an independent study (Blalock, Cushman & Nelson, 2009) using nine students who had not yet completed their initial polygraph training course, mean accuracy was sufficient to meet the high Marin certification standard of 86% or better.

All of the cases used in the Nelson *et al.* (2008) and Blalock *et al.* (2009) analyses were conducted using the Federal Zone Comparison Technique (FZCT), a method with three relevant questions. A second version of the ZCT uses two relevant questions, and for which the ESS has not previously been tested. The purpose of the present analysis was to determine whether the ESS would generalize to the two-question version of the ZCT.

In 2008 an unrelated study was reported in which 100 two-question confirmed ZCT cases were analyzed (Meiron, Krapohl & Ashkenazi, 2008), and those scoring data were reused here to test the ESS. There were two sets of scorers in the Meiron *et al.* project, three Israeli Government examiners[2] and three US government examiners. The ESS was applied to each of these data sets. As a benchmark against which to compare the ESS, the decisions of US Government examiners using standard scoring methods as reported in the Meiron *et al* study are included.

The original 7-position scores in the Meiron data were collapsed to 3-position scores. For the pneumograph and cardiograph channels, the scores became -1, 0 or +1. For the electrodermal channel the scores were -2, 0 or +2. The scores for each case were summed by spot and total, and the two-stage (Senter, 2003) ESS decision rules were imposed. First, if the total score was +2 or greater the decision was NDI. A total score of -4 or lower the decision was DI. If the results would be inconclusive, the second stage is initiated. In the second stage, any spot score of -6 or lower would result in a DI call[3] (Nelson, 2010). All else is inconclusive. See Table 1.

[1] This article is one in a series titled Best Practices. The opinions in this article are those of the author, and do not necessarily represent those of the US Department of Defense or Government. The author wishes to express appreciation to Mark Handler for his suggestions to an earlier version of this paper, and to Ray Nelson for computational assistance.

[2] The Israeli government scorers actually scored the data twice, once using the Backster "Either-Or" Rule and a second time where this rule was not used. A principal aim of the ESS is simplicity, and in that vein the simpler of the two data sets (without the "Either-Or" Rule) was used for the present project. Though not found in Table 1, there were no statistical differences for the ESS with the "Either-Or" data set when compared to the others.

[3] Note: For the three-question ZCT the spot score threshold is -7. The total-score thresholds are the same for both the two-question and three-question versions of the ZCT.

Table 1. Percentages of average correct, erroneous, and inconclusive outcomes (and confidence intervals) for groups of scorers, and the average percentage of decision agreement (DI, NDI or inconclusive) among pairs of scorers for the three groups. There were no statistical differences for any of the percentages in each decision category among the three groups.

	Correct	Error	Inc	Correct w/o Inc	Average Agreement
Standard US Federal	85.7 (78.7 - 92.4)	5.3 (1.0 - 9.8)	9.0 (3.6 - 14.6)	94.1 (89.2 - 98.9)	83.3 (77.5 - 89.2)
Israeli scorers w/ ESS	83.3 (79.0 - 92.5)	6.0 (2.1 - 11.6)	10.7 (2.2 - 12.6)	93.3 (87.5 - 97.7)	86.7 (82.4 - 92.3)
US Scorers w/ ESS	86.0 (76.1 - 90.3)	6.7 (1.5 - 10.5)	7.3 (4.7 - 16.8)	92.8 (88.2 - 98.3)	87.3 (81.6 - 92.0)

Conclusion

These data add further support for the ESS, showing performance statistically equivalent to the system most commonly used in the field. Combined with the results of the Blalock *et al.* (2010) study indicating high accuracy even with low field experience, and the increase in reliability that arises from fewer scoring rules, the ESS may be advantageous in settings where standardization is needed most.

Limitations

These findings allow comparisons among the three sets of scores, but not for the overall validity of either ESS or the two-question ZCT due to potential anomalies in case selection (See Meiron, Krapohl & Ashkenazi, 2008). Moreover, the scoring features used by the original scorers correspond with, but are not identical to, the three features on which the ESS was developed. Consequently, some variance in individual scores is expected between the ESS and the systems that have more or different features.

References

ASTM (2005). *E2324-04 Standard Guide for PDD Paired Testing*. ASTM International: West Conshohocken, PA

Blalock, B., Cushman, B. and Nelson, R. (2009. A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281-288.

Meiron, E., Krapohl, D.J., and Ashkenazi, T. (2008). An assessment of the Backster "Either-Or" Rule in polygraph scoring. *Polygraph*, 37(4), 240-249.

Nelson, R. (2010, Feb). *Empirical Scoring System*. Presentation to the New Mexico Polygraph Association.

Nelson, R., Krapohl, D.J., and Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph*, 37(3), 185-215.

Senter, S.M. (2003). Modified general question test decision rule exploration. *Polygraph*, 32(4), 251-263.