APA Research Committee Report: Proposed Usage for an Event-specific AFMGQT Test Format[1]

Raymond Nelson, Mark Handler, Marty Oelrich and Barry Cushman

Abstract

The members of the APA research committee respond to an inquiry about modifications to the AFMGQT format for event-specific examinations. With regard to the generalizability of presently available scientific knowledge regarding event-specific test formats with two or three relevant questions, we reviewed the published evidence on the validity of the super-dampening hypothesis, and the effectiveness of outside-issue or "symptomatic" questions to reduce inconclusive results, increase test accuracy, or detect the presence of outside issues. We considered whether the inclusion of outside-issue questions affects either the distributions of test scores or the criterion accuracy of test results. Empirical evidence regarding the effectiveness of outside-issue questions is confounded and therefore uninformative. Research on outside-issue questions and the super-dampening hypothesis has failed to conclusively demonstrate the hypothesized effects, and test question formats that do not include outside-issue questions have been shown to produce mean accuracy rates that equal or exceed that of test question formats that do include these questions. No basis of evidence was found to support a conclusion that our present knowledge and normative data cannot be generalized to the AFMGQT if used as an event-specific test format with two or three relevant questions.

Members of the Research Committee of the American Polygraph Association (APA) have received questions about the validity of the use of the Air-Force Modified General Question Technique (AFMGQT) V1 and V2 question formats (Department of Defense, 2006a) as an event-specific single-issue question format. Specifically, these questions have focused on the potential for the use of the AFMGQT question formats in a manner similar to the Federal You-Phase (You-Phase) format and Federal Zone Comparison Test (ZCT) format (Department of Defense, 2006a). These questions pertain to the generalizability of the presently available normative data and published evidence of criterion validity. The inquiries addressed whether our present knowledge is sufficient to quantify the degree of uncertainty and confidence level for test results obtained using the AFMGQT format as an event specific single-issue examination format. In other words, would our normative data generalize from a ZCT to an AFMGQT format?

The proposed use of the AFMGQT includes two (2) relevant questions (RQs) bracketed in sequence with three (3) comparison questions (CQs), (i.e., CQ, RQ, CQ, RQ, CQ). The format also includes non-diagnostic questions in the form of one neutral and one sacrifice-relevant question. The two RQs of the proposed format would be formulated to address the details of a single issue of concern in a manner similar to the target selection and question formulation practices common to the Federal You-Phase technique. The You-Phase uses a similar sequence of scored questions, with the addition of two outside-issue questions.

[1] This report is the product of the APA Research Committee. It was submitted to the APA Board of Directors, which accepted it. It should be considered a research report, and does not represent an official position of the American Polygraph Association.

A similar proposed sequence has been suggested using the AFMGQT format in event-specific testing contexts with three (3) RQs in sequence with three (3) CQs, (i.e.,CQ, RQ, CQ, RQ, CQ, RQ). The proposed target selection and question formulation approach would be similar to that of the ZCT, a test question sequence that can include both primary (i.e., direct involvement), secondary (i.e., indirect involvement) relevant questions, and procedural questions consisting of one neutral and one sacrifice-relevant question. Test questions would be selected and formulated using the principles commonly employed with the ZCT format with the exception of the exclusion of the outside-issue questions.

The You-Phase and ZCT formats are used in event-specific diagnostic contexts. They are also used in single-issue screening contexts, as a stand-alone single-issue screening format, or as a single-issue break-down test for use following an unresolved multiple-issue screening exam. Whether employed in a single-issue screening context or event-specific investigation context, test questions of You-Phase and ZCT formats are interpreted with the assumption of non-independent criterion variance (Department of Defense, 2006b). That means behavior that affects the criterion state (i.e., guilt or innocence) of any individual target question may also affect the criterion state of other target questions. Assumptions about independent and non-independent variance are expressed in field practice through the selection of decision rules and normative data that are used to parse the examination result and determine the level of statistical significance.

Traditional usage of the AFMGQT has been as a multi-facet format for the investigation of known or alleged problems, and as a multi-issue screening format. Scoring and interpretation of AFMGQT test data has been traditionally accomplished with the assumption that the criterion state of each target question varies independently (American Polygraph Association, 2001; Department of Defense, 2006b). This means an assumption that the criterion state of individual target questions is distinct from the criterion states of other examination targets (innocent and guilty). An important difference in the proposed use of the AFMGQT format is the selection and formulation of AFMGQT target questions for which the criterion state of the target stimuli would be assumed to vary non-independently and thus evaluated using grand total scoring.

Research has not specifically addressed this particular usage of the AFMGQT, so scientific questions about criterion accuracy and normative data depend on an evaluation of our knowledge regarding existing event-specific examination formats. Is it scientifically reasonable and responsible to generalize knowledge from the You-Phase and ZCT formats to the proposed use of the AFMGQT format? Polygraph examination results may play a role in decisions that might affect community safety, individual liberties, future opportunities, and information security. Any conclusion regarding these matters will be best informed by a careful consideration of the scientific evidence and not by mere speculation. The practical importance of this inquiry involves whether field practitioners who use this modification are compliant with expectations for evidence-based practices and the use of validated polygraph techniques. If not, members of the public may have a reasonable expectation for notification of its use as an experimental technique as a component of informed consent and authorization for testing.

The APA Research Committee takes the position the answer to questions about the validity of the proposed use of the AFMGQT exists in the evidence pertaining to two fundamental issues: 1) whether outside issue questions increase the effectiveness of the polygraph test, and 2) whether presently available knowledge about criterion accuracy and the normative distributions of scores of innocent and guilty persons is generalizable and sufficient to quantify the degree of uncertainty associated with test results.

Discussion

At the core of the present inquiry is a fundamental issue regarding whether our present knowledge is sufficient to allow us to generalize the degree of uncertainty and confidence level to the expected range of accuracy for You-Phase and ZCT exams. The

answer to this question will be best determined by the presence or absence of meaningful differences between the proposed use of the AFMGQT formats and the more traditional You-Phase and ZCT formats. A cautious view would hold that the existence of any meaningful difference would require additional study before proceeding to endorse the proposed use of the AFMGQT. In this case, the notion of "meaningful" differences must be defined as any difference for which evidence has been shown to affect either the distribution of numerical test scores or the criterion accuracy of test results. Additionally, any clearly articulated hypothesis defining a plausible mechanism by which numerical test scores or criterion accuracy would be expected to differ could be a basis for caution and additional study before proceeding to endorse the proposed use of the AFMGQT.

The proposed use of the AFMGQT as an event-specific test format hinges on the question of whether available normative data for You-Phase and ZCT exams can be generalized to the AFMGQT if the target selection, question formulation, and test data analytic methods are similar for formats with two or three relevant questions. Normative data are used to calculate the degree of uncertainty associated with a test result, and to calculate decision cutscores that determine the level of statistical significance of categorical conclusions based on probabilistic test data. Normative data exist for both the Federal You-Phase format and the Federal Zone Comparison Test formats, including normative statistics for 7-position scores (American Polygraph Association, 2011) and Empirical Scoring System (ESS) scores (American Polygraph Association, 2011; Nelson et al., 2011) based on an assumption of non-independent criterion variance among the target questions.

Some differences exist in the traditional use of the AFMGQT and the You-Phase and ZCT formats. Among these differences are the approach to target selection and question formulation, and assumptions about independent or non-independent criterion variance when scoring the examination data. Normative data are available for the AFMGQT format when the test stimulus questions are scored and interpreted with the assumption of independent criterion variance (American Polygraph Association, 2011). The proposed use would eliminate this important difference.[2] Differences would remain in the numerical position of each question in the sequence,[3] though there is no plausible rationale suggesting that any differences in question numbers or the proposed changes in location could affect test performance characteristics.

One other difference between the proposed use of the AFMGQT format and the traditional use of You-Phase and ZCT formats is the exclusion of outside-issue questions, referred to by examiners as "symptomatic" questions. Two examples of outside-issue questions can be found in publications by the Department of Defense (2006), including: 1) "Do you believe I will only ask you the questions we reviewed," and 2) "Is there something else you are afraid I will ask you a question about?"[4]

Exclusion of outside-issue questions from the proposed AFMGQT event-specific test question sequences deserves further discussion. Inclusion of outside-issue questions in the test question sequence is the

[2] The proposed use of the AFMGQT would not prevent the continued use of the AFMGQT in the traditional manner when circumstances indicate the assumption of independent criterion variance to be useful.

[3] Relevant questions are located at position 5 and 7 for You-Phase formats and positions 5, 7 and 10 for ZCT formats, whereas these questions would be located at positions 4, 6 and 8 for the proposed use of the AFMGQT.

[4] Other versions of outside issue questions have also been described.

practical expression of the incorporation of the super-dampening hypothesis[5] into the comparison question model.[6]

The super-dampening hypothesis suggests that issues outside the scope of the investigation target(s), if they are of greater concern to the examinee than the investigation target questions, are a cause of inconclusive results (Backster, 1962). Exclusion of outside-issue questions represents an exclusion of the super-dampening hypothesis from the model. A corollary to the super-dampening hypothesis is that the use of outside-issue test questions will enable an examiner to detect the existence of an outside issue. Implicit in every hypothesis is the expectation of increased effectiveness of the model. In the present investigation, evidence of increased model effectiveness must be observed in terms of some dimension of criterion accuracy (e.g., sensitivity, specificity, error rates, or overall decision accuracy) as a result of a decrease in the rate of inconclusive results. Additionally, implicit in the inclusion of any hypothesis is a description of how to operationalize and quantify any anticipated effect.

A hypothesis is valuable, and therefore acceptable, if it offers some increase in the effectiveness of the model to which it is applied. False hypotheses are of no value and must be either discarded and replaced or subjected to modification until a new hypothesis is supported by replicable evidence. This is an inherent aspect of the scientific method. The requirement for acceptance of a hypothesis is simple, though not necessarily easy to achieve: replicable evidence must exist to show that the observed phenomena or data actually function as expected. In the present case, the super-dampening hypothesis purports to explain the occurrence of inconclusive test results, and the related outside-issue questions purport to either detect or reduce the interference that outside issues may impose on the test data and test result. The overarching measure of the validity of the super-dampening hypothesis and outside-issue questions is whether the criterion accuracy of polygraph test results is improved in terms of discrimination accuracy, (i.e., test sensitivity and test specificity) or errors (i.e., false-positive and false-negative errors). A related issue is whether incorporation of the

[5] The super-dampening hypothesis is referred to by Backster (2001) as the "super-dampening concept." The term "concept" is not found in general usage in this manner in the scientific milieu. For any idea, explanation, or suggestion to merit consideration as a scientific hypothesis it must first be stated in a way that is falsifiable and testable – which requires some form of objective measurement, or minimally requires a movement towards objective measurement in the form of published clinical observations that can be easily replicated. Hypotheses that are offered in a manner that cannot be tested, and therefore cannot be falsified, are viewed as flawed hypotheses which are therefore false hypotheses – else we risk engaging in pseudo-scientific practices. Hypotheses are experimental questions or research questions, which must be regarded with caution until there is replicated evidence to support their validity. Scientific theories are those working hypotheses for which replicated evidence exists to support the idea or explanation. Theories are therefore hypotheses for which evidence suggests they are correct and are not inconsistent with other evidence. Because it is not possible to know everything, scientists are obligated to continue learning. The role of science is to continue to study our ideas and to compare our present knowledge and assumptions to an ever-expansive database. Curious probing of the limits of knowledge and the accuracy of assumptions will eventually result in the observation of evidence that is inconsistent with or not explained by current working theories – at which time the obligation is to either modify the hypothesis in a manner that can account for both old and new evidence, or simply discard the hypothesis as incorrect and begin again with a new hypothesis to explain both the old and new evidence. Laws of science are those ideas for which the evidence and replication is so strong and so robust that it is expected that no new evidence will ever controvert their validity. The super-dampening "concept" is, in actuality, the super-dampening hypothesis.

[6] Use of this term conveys that polygraph testing does not test lies per se, but instead tests for statistical significance or error by recording and scoring proxy data, in the form of differences in strength of physiological responses as indicative of the salience of different types of test stimuli that are combined in mathematically optimal ways to represent or "model" the act of deception or truthfulness within a high range of accuracy that is significantly greater than chance. The objective of scientific testing – when direct observation or measurement of the phenomena of interest is not possible – is to quantify the degree of uncertainty associated with a conclusion.

super-dampening hypothesis and outside-issue questions into the test structure can be expected to account for a causal change in the distributions of scores of guilty or innocent persons.

If the super-dampening hypothesis is a true or correct hypothesis, then rejection of the super-dampening hypothesis and exclusion of outside-issue questions from the test structure might be expected to adversely affect the distribution of numerical scores and criterion accuracy of test results. This would then affect the generalizability of our knowledge regarding the normative distributions of scores of guilty and innocent persons to test formats that do not include outside-issue questions. If the super-dampening hypothesis is false or incorrect (i.e., outside-issue questions have no effect on test scores or test accuracy) then the exclusion of outside-issue questions from the test question sequence can be expected to have no effect on the distributions of numerical scores and criterion accuracy of test results. It is therefore critical to understand the basis of scientific evidence for the super-dampening hypothesis and outside-issue questions.

Review of scientific literature on the super-dampening hypothesis.

Capps, Knill and Evans (1993) conducted a study of outside-issue questions using three experienced examiners who were licensed in their respective states and members of the American Polygraph Association. The examiners, who averaged over 5 years' experience, agreed to conduct every other examination using a ZCT format while including the outside-issue questions or substituting them with neutral questions. A total of N = 150 field examinations were collected, including 75 ZCT exams with outside-issue questions and 75 ZCT exams without outside-issue questions. Data were evaluated by the original examiners who reported a total of 12 inconclusive cases when outside-issue questions were excluded and four (4) inconclusive results when outside-issue questions were included.

Capps, Knill and Evans (1993) reported the results of a goodness of fit chi square (X^2) test suggesting a statistically significant difference (X^2 [1] = 4.48, p = .034). However, they did not report what subjective or unquantified methods the original examiners used to decrease their inconclusive results when outside-issue questions were included. Results of the original examiners are confounded by the fact that the examiners may have relied on some extrapolygraphic information, in the form of interpersonal observations or case background information, in the formulation of their conclusions. The exact cause of the reduction of inconclusive questions cannot be conclusively attributed to the asking of the outside-issue questions. Capps, Knill and Evans also reported the results using an automated statistical algorithm, which resulted in 8 inconclusive results when outside-issue questions were excluded and 5 inconclusive results when outside-issue questions were included, but they did not report the results of a statistical test of the results using the objective measurements from the automated scoring algorithm.

Honts, Amato and Gordon (2001) reanalyzed the field data reported by Capps, Knill and Evans (1993) and showed that the outside-issue questions accounted for about 3% of the variability in conclusive vs. inconclusive results. Because Capps, Knill and Evans did not report the results of a statistical test of the automated examination results, Krapohl and Ryan (2001) re-evaluated the reported results and calculated the chi square statistic using the same 2x2 contingency method which showed that the difference was not statistically significant (X^2 [1] = 0.758, p = .384), indicating that differences in inconclusive rates when outside-issue questions were included or excluded did not differ from chance. An important difference between the automated results and the manual scores is that the automated scoring method could not make any use of extrapolygraphic information, and relied only on recorded physiological responses. It is also possible that the participant examiners may have known or guessed the purpose of the experiment, and that the observed results were influenced by expectations. Although the results of Capps, Knill and Evans are of interest, confounded results cannot be accepted as a sufficient basis for a conclusion that the outside-issue questions actually provide the hypothesized advantages.

Results of the Capps, Knill and Evans (1993) study were not sufficient to reject the null hypothesis that the inclusion of outside-issue questions makes no significant contribution to the reduction of inconclusive results based on recorded physiological responses. Neither could the Capps, Knill and Evans study reject the null-hypothesis regarding any ability to detect the actual presence or absence of an outside-issue. Importantly, they did not investigate the actual presence or absence of an outside-issue and thus could not investigate whether outside-issue questions contributed to any ability to discriminate the presence or absence of outside issues. Lastly, the sampling design prevented any ability to investigate whether the super-dampening hypothesis and outside-issue questions increased the criterion accuracy of polygraph test results or distribution of truthful and deceptive scores.

Honts, Amato and Gordon (2001) studied the outside-issue question using a factorial design involving programmed innocent and guilty persons, existence or non-existence of an outside issue, and the inclusion or exclusion of outside issue questions from the test question sequence. There were N = 92 participants randomly assigned to the eight different conditions. This research was designed to investigate the effect of the super-dampening hypothesis and outside-issue questions on the criterion accuracy of the test, and whether differences exist in the way that guilty and innocent persons respond to outside-issue questions. Also addressed was whether experienced Federal examiners could use the outside-issue questions to detect the existence of outside issues, and whether the outside-issue questions could be subjected to a formal analysis. Data were analyzed by three instructors from the Department of Defense Polygraph Institute (now called the National Center for Credibility Assessment).

Honts, Amato and Gordon (2001) reported that the existence of an outside-issue had little effect on the test results of guilty persons, but was associated with more negative test scores for innocent participants. Scorers had no ability to discriminate the existence of outside issues at rates greater than chance. The investigators calculated a decision efficiency coefficient (Kircher, Horowitz & Raskin, 1988) to determine the effectiveness of the outside issue question at detecting outside issues, but were unable to develop a formal statistical or numerical solution to effectively score or interpret the outside-issue questions. Honts, Amato and Gordon reported that a single statistically significant discriminate function was identified in the form of a difference between responses to outside-issue questions and relevant questions. They reported a canonical correlation of .215 with classification accuracy of .667 for outside-issue absent participants and .645 for outside-issue present participants, and reported they did not expect the results to survive cross-validation.

Honts, Amato and Gordon (2001) reported that the outside-issue questions did not contribute to increased criterion accuracy or increased detection of the presence of outside issues. Results of a multivariate ANOVA indicated that outside issue questions could be used as valid comparison questions, though there is no previous or subsequent precedent for this in field testing protocols. Results from the Honts, Amato and Gordon study did not support the super-dampening hypothesis and wrote that “concerns that undiscovered crimes may overwhelm responses to relevant test questions appear to be groundless” (p. 73).

A study by Krapohl and Ryan (2001) involved N = 100 field examinations that were selected from an archive of confirmed cases at the Department of Defense and subjected to 7-position blind numerical scoring by an experienced examiner who also made subjective notation about the presence or absence of outside issues for each recorded channel and each question. Data were also evaluated using an automated rank order method applied to relevant, comparison and outside-issue questions. Krapohl and Ryan reported virtually no correlation between the number of reactions to outside-issue questions and the total numerical scores (r = .07), and virtually no correlation between the rank difference scores (i.e., difference in summed ranks of relevant and comparison questions) and the rank scores of outside-issue questions (r = .06). Similarly, there was no correlation between the rank difference

scores and outside-issue questions for innocent cases (r = .04). A slightly stronger correlation was found between the rank difference scores and outside-issue questions for guilty cases (r = -.18),[7] but the relationship was not statistically significant. There was no evidence of dampening of numerical scores as a function of responses to outside-issue questions and no support for the super-dampening hypothesis. Krapohl and Ryan concluded that their analysis controverted the super-dampening hypothesis.

Conclusion

This review of the published literature revealed that outside-issue questions have been shown to be ineffective at achieving the objective implied by the language of these questions (i.e., testing for the presence of outside issues that might explain or reduce the occurrence of an inconclusive examination result). We found no evidence to support a conclusion that numerical scores or test accuracy vary as a function of responses to outside-issue questions. We therefore conclude that our present knowledge regarding criterion accuracy and normative distributions of scores can be generalized to other named test formats if the test target selection and question formulation, stimulus presentation, and test data analysis protocols are similar. We further conclude that our present knowledge-base is sufficient to quantify the degree of uncertainty of results achieved with the proposed use of the AFMGQT within the range of effectiveness shown for other existing event-specific polygraph techniques with two or three relevant questions, regardless of the inclusion or exclusion of un-scored outside issue questions.

Acceptance of the super-dampening hypothesis, and endorsement of assumptions that outside-issue questions play a causal role in inconclusive results, test accuracy, or the distribution of test scores requires un-confounded and replicable evidence to support the rejection of the null hypothesis of no effect. However, evidence in support of the super-dampening hypothesis and outside-issue questions is virtually non-existent at this time. At the present time, the basis of published evidence amounts to a small number of studies that collectively fail to support the validity of the super-dampening hypothesis or the use of outside-issue questions to achieve the hypothesized effect. We also note that there has been some change or shifting of the hypothesized effects, and that none of the suggested benefits are supported by empirical evidence.

At the present time there is no published method for the quantitative analysis of outside-issue questions and no published normative data with which to attempt to quantify the degree of uncertainty regarding any interpretation of the meaning of responses to these questions. Similarly, there is no published description of the efficacy of any structured clinical procedure or qualitative method for the interpretation of responses to these questions. Interpretation of responses to outside-issue questions remains an unquantified and unstructured clinical process. Although some evidence suggests that the existence of outside issues may be related to more negative scores for innocent persons, it is not known whether this is a function of increased responding to relevant questions or decreased responding to comparison questions. Regardless, we found no basis for any conclusion that generalization of our knowledge of criterion accuracy and normative distributions of scores of guilty and innocent persons should be regarded as a function of the inclusion or exclusion of the outside-issue questions themselves.

Some controversy and discussion followed previous publications on the outside-issue question, primarily in response to disagreements expressed by the originators and proponents of the super-dampening hypothesis. These disagreements offered no additional evidence to the discussion, and

[7] This is somewhat consistent with the finding of Honts, Amato and Gordon (2001) that outside-issue questions may function as a form of comparison question.

rested upon personal opinion and conjecture. If the past is indicative of future potential, some objection can be expected to these conclusions. If such objections are accompanied by un-confounded evidence to support them, then they should be taken seriously and this matter should be reconsidered. Objections or arguments offered without evidence are a form of hypothesis in need of study. There is no benefit to the profession to ignore available evidence in favor of unconfirmed or untested hypothesis, as this form of confirmation bias serves only to gratify a secondary objective while interfering with scientific integrity.

Examples of past objections to the scientific evidence on the super-dampening hypothesis and outside-issue questions can be found in the published record. Backster (2001) objected to the use of "pooled" raw data, and suggested that each test chart is "unique" in its ability to predict the examiner's success. In this usage "unique" requires definition – as it appears distinct from the scientific and statistical issue of independence which refers to the degree of covariance between items (charts) that are expected to covary. Polygraph test charts within a single exam are non-independent in that they include the same stimuli addressing the same stimulus targets with the same examinee. If the stimulus questions on one chart are affected by past behavior then it stands to reason that the same past behavior can be expected to influence reactions on subsequent test charts.

Backster (2001) suggested using only those cases for which there was a response to outside issue questions and no response to relevant questions and comparison questions. To do this would be a serious breach of research protocol as it would amount to begging the questions or "cherry-picking" the data to conform to a prior conclusion. It would be the scientific equivalent of writing a story to fit a desired ending. Backster further suggested that the absence of evidence is no basis for excluding outside issue questions from the test structure and administration, and offered the analogy that removal of outside-issue questions from the question format due to the absence of evidence would be the equivalent of removing smoke detectors because of the absence of fire. To this we point out that the analogy fails because smoke detectors have been shown to be capable of detecting fires, whereas outside-issue questions have not been shown to function as hypothesized. More fundamentally, it is important to adhere to the basic principles of ethical scientific testing and evidence-based practices: evidence should normally be required before including an experimental hypothesis into field practices that affect both individuals and the public. We note that discussions about evidence-based practice had not begun to be clarified within the polygraph profession at the time of Backster´s work on the super-dampening hypothesis and the outside issue questions.

Backster (2001) began to re-frame the purpose of the outside-issue questions when he wrote the following: "Again, these questions are designed to indicate the success of the examiner in gaining, at a minimum, the very limited trust of the examinee that no unreviewed questions will be asked" (p. 213). Complex assumptions such as the ability to test the examinee´s trust in the examiner should normally require evidence and replication before being incorporated into procedures that affect individuals and communities. The suggestion that outside-issue questions provide evidence of the level of trust which the examinee holds for the examiner, or that the degree of trust is expressed in inconclusive results, is in essence a highly speculative and largely untestable, and therefore unscientific, hypothesis within the polygraph testing milieu.

No evidence has been offered to support the notion that outside-issue questions can indeed test for the "trust of the examinee." Similarly, no description exists regarding the psychological basis for an assumption that quality or "trust," as an attribute of the rapport between the examinee and examiner, is manifested in the same autonomic responses that are used to calculate the normative level of statistical significance of the differential loading of attention and responses onto the comparison and relevant test stimuli. Honts, Amato and Gordon (2001) wrote that asking "'Is there something else you are afraid I will ask you

about?' is the functional equivalent of 'Prior to 1998 did you ever do anything that was dishonest or illegal?'" A more plausible and testable hypothesis than the trust-hypothesis would suggest that observed physiological responses function as a combination of cognitive and emotional loading related to the goal of passing the polygraph test and past behavior that is referenced by the test stimulus questions. This form of hypothesis can be more easily studied in the context of differential responses to different types of test stimuli, and can more readily account for previous known phenomena and observations than asking questions about outside issues may stimulate responses to outside issues. Additional research is needed in this area.

In 1962, Backster had not yet introduced the trust-hypothesis, and described the super-dampening hypothesis as a dampening of all responses that would normally occur:

> With such a person the outside issue, about which he is so apprehensive, is much more important to him – or more directly affects his well-being – than does the reason for the polygraph examination, thus causing a 'super-dampening' of all responses that would ordinarily have occurred, including the dampening of all response to review stimulation questions asked the innocent suspect (p. 65).

Embedded in this statement are a number of esoteric assumptions and corollaries to the super-dampening hypothesis, some of which have subsequently been met with controverting evidence. The most central of these is the notion of apprehension or fear about well-being as the basis of responses to relevant and comparison stimuli – a hypothesis that has been shown repeatedly to be false through the high criterion accuracy of polygraph techniques that make use of non-manipulative directed-lie comparison questions.[8]

Other objections have been expressed by Matte (2001a; 2001b) who suggested that Krapohl and Ryan (2001) had failed to recognize the limitations of the Capps, Knill and Evans sample size (N = 150). It is highly unlikely that the sample size was overlooked, as Krapohl and Ryan commented on the sample size in their published report. Additionally it would have been impossible to overlook the sample size when calculating the chi square statistic that Matte criticized despite the fact that Capps, Knill and Evans (1993) used the same statistic. Perhaps a different statistical method would be selected today – possibly a robust statistical method such as bootstrapping, which does not depend on assumptions about the shape of the distribution of data, or a Poisson statistic that is useful with infrequent events – but the chi-square statistic was most likely the preferred method at the time of the Capps, Knill and Evans study, and there is some sense in using the same method to calculate the missing statistic.

Matte (2001b) argued that the confounded Capps, Knill and Evans (1993) results as adequate to reject the null hypothesis, and also re-framed the purpose of the outside-issue questions, around the hypothesized effects of providing reassurance and discrimination regarding trust and confidence, when he wrote the following:

> To begin, it should be understood that the review and subsequent incorporation of the symptomatic questions into the Zone Comparison Technique is designed primarily to reassure the examinee that no unreviewed questions will be asked during the administration of the polygraph test during which the physiological data is collected, and secondarily to identify the examinee who is not convinced that no unreviewed question will be asked during the test. (p. 220)

[8] A summary of studies on polygraph techniques using directed and probable lie comparison questions was published by the American Polygraph Association, 2011.

In addition to the fact that the assurance hypothesis was attached to the outside-issue questions only after the emergence of evidence to controvert the effectiveness of these questions, the psychological basis for the hypothesized assurance effect cannot be determined. Matte also suggested the use of outside-issue questions as the last question in the series, and offered the rationale that examinees may "relieve"[9] on the last question in anticipation of the end of the test. This represents a substantial departure from earlier suggested purpose and rationales for outside-issue questions. Regardless, the findings of Capps, Knill and Evans remain confounded and insufficient to support a conclusion regarding the validity of the outside issue questions. A scientific approach requires that we not simply reject uncomfortable information, and not simply focus on flattering evidence.

In the 13 years since the publication of Honts, Amato and Gordon (2001) and Krapohl and Ryan (2001), no new attempt has been made by proponents of the outside issues questions to resolve the confounded results of Capps, Knill and Evans (1993). No evidence has been added to our scientific knowledge-base regarding the super-dampening hypothesis and outside-issue questions since that time.

In response to any potential suggestion that outside-issue questions are expected not to test for the influence of an outside-issue but to correct for such influence, we note that there is no rationale in the published literature in psychology, physiology, psychophysiology, scientific testing, or statistical decision theory suggesting that the presentation of a test stimuli can be expected to correct for a problem that cannot be reliably quantified. A more reasoned approach would be to reinforce the general principle that examiners should endeavor to conduct the interview and data acquisition in a manner such that the examinee is not distracted. Honts, Amato and Gordon (2001) addressed this in their observation that outside issues may be related to false-positive errors, and that this may be exacerbated by the use of ineffective examination formats or accusatory interviewing approaches that may increase the level of distraction due to outside issues. They suggested further research in this area. Krapohl and Ryan (2001) noted that outside-issue questions may introduce distraction to the testing contact when they discussed reported concerns that outside-issue questions raise suspicions among some persons. They also pointed out that the inclusion of outside-issue questions changes an explicitly single-issue exam into an explicitly multiple-issue exam.

In consideration of the fact that the validated underlying principles for the proposed use of the AFMGQT format are identical to those of other event-specific formats, and the evidence suggesting that outside-issue questions and related assumptions to be a false hypothesis, the only remaining potential difference between the proposed use of the AFMGQT and other event-specific test formats involves the hypothesis that that the mere inclusion of outside-issue questions may prompt some indescribable fundamental changes in the function or effectiveness of the test target stimuli. At the present time there is neither any evidence nor any plausible rationale to support this speculation, which must, like other hypotheses, be reasonably supported by some description of the psychological or physiological mechanism by which some change can be expected to occur, and must eventually be supported by evidence before being accepted. With the exception of outlier results that should be regarded with great concern (American Polygraph Association, 2011), available evidence does not support the notion that test question formats which include outside-issue questions outperform formats without outside-issue questions.

[9] The construct of "relieve" in this usage is not found in the psychophysiological literature outside of discussions of the Backster polygraph techniques. It is not clear whether this term refers to a form of reaction due to relief or a return to tonic level.

Summary

It is the position of the APA Research Committee that criterion validity and generalizability of knowledge pertaining to testing formats of all types depends not on the name of the question format, and also not on the incorporation of false hypotheses into the test structure. Criterion validity and generalization depends, in part, on the validity of the underlying constructs that determine the production, recording and mathematical or statistical quantification of the recorded data. The central underlying construct for comparison question test formats of all types, including the AFMGQT, Federal You-Phase, and Federal Zone Comparison Test formats, is that the strength of response to different types of test stimuli will vary at statistically significant levels as a function of the guilt or innocence state of the examinee with respect to a behavioral target issue, and that differential salience and response magnitude can be quantified to predict the criterion state of the examinee within a known range of accuracy that is significantly greater than chance.

Event-specific examinations with two and three relevant questions have been shown to provide criterion accuracy at rates greater than chance (American Polygraph Association, 2011), but evidence at this time suggests that these formats all work in spite of, not because of, the false or unproven super-dampening hypothesis and outside-issue questions. Although these questions are intended to test for the possible role of an outside-issue as a cause for inconclusive results of numerically scored exams, interpretation of responses to outside-issue questions appears to be a phenomenological concern for which there are no published rules and no evidence to support the effectiveness of the clinical use of these questions. In summary, use of outside-issue questions remains a hypothetical suggestion for which the evidence is presently lacking and not likely to emerge. If one were to argue that the mere inclusion of the outside-issue questions somehow increases validity, even if the validity of these questions themselves is questionable,[10] then ethical requirements for evidence-based practice dictates that this too should require replicable evidence before proceeding to require the inclusion of these questions into the question format.

Calculation of the degree of uncertainty or level of statistical significance of summative examination scores is a function of two main factors: 1) whether the test target questions are interpreted with the assumption of independent criterion variance (i.e., past behavior that affects the guilt vs. innocence criterion of any one question is thought to have no effect on the guilt vs. innocence criterion of other target questions), or with the assumption of non-independent criterion variance (i.e., past behavior that affects the guilt vs. innocence criterion of any one question is expected to potentially affect the guilt vs. innocence criterion of other target questions), and 2) the number of stimulus target questions, which together with the number of repetitions of the test question sequence will affect the total number of stimulus-response observations that compose the test data-set, which may affect the applicability of available normative data to the summative test scores.[11]

It is the conclusion of the research committee that the proposed use of the AFMGQT conforms sufficiently to the same validated principles as the test question formats for the Federal You-Phase and Federal Zone Comparison Test formats, and that generalization of available normative data appears reasonable. Similarly, with consideration for the lack of any evidence, description, or indication that outside-issue questions play a role in the formulation of test scores, generalization of presently available

[10] We note that the rationale is unclear as to why the inclusion of a false-hypotheses would be expected to increase criterion accuracy.

[11] Numerical transformation and aggregation models that do not involve the summation of test data (i.e., rely on averaging, standard scores or other statistical functions) may be less affected by the number of test questions or number of repetitions.

knowledge regarding test accuracy also appears to be a reasonable suggestion. Although the presence of outside issues may be related to test scores, there is no evidence to support the hypothesis that responses to outside-issue questions themselves have any effect on numerical scores, test accuracy, or the generalizability of presently available normative data used to calculate the level of statistical significance of an examination result.

We note that the proposed use of the AFMGQT adds no new hypothesis to the testing model, but instead removes questions related to a hypothesis that appears to be false. Although there is no direct experimental research regarding the effect of inclusion or exclusion of outside-issue questions in the test question sequence, a comparison of published evidence on the criterion accuracy rates of different test question formats shows that inclusion or exclusion of outside issue questions results in numerical score distributions for which the normal ranges substantially overlap (American Polygraph Association, 2011). Indeed, test question formats that attempt no use or interpretation of outside-issue questions have been shown to produce levels of criterion accuracy that equal or exceed the criterion accuracy of formats that include the traditional use of outside issue questions.

Several cautionary observations are worth noting, though they may seem obvious once they are articulated. First, it does not serve the profession to impose the use of a false hypothesis or to restrict the use of test formats that do not incorporate a false hypothesis. Additionally, it does not serve the profession to make examiners vulnerable to criticism around false hypotheses and procedural requirements that make no difference. Nor does it serve the profession or the community to perpetuate rules that are not additive to the effectiveness of the polygraph testing model. Neither does it serve the profession nor the community to create or impose standards that would please individual personalities while neglecting the obligation to formulate meaningful evidence-based practice standards. It does not increase our knowledge or intelligence to direct our attention and resources to issues that do not actually contribute to test effectiveness. Nor does improve the standing or stature of the polygraph profession in the broader realm of science, testing, and forensics to remain anchored to concepts and terminology that are not supported by evidence, or are inconsistent with science, and which do not contribute to increased accuracy of discrimination between deception and truthfulness. Scientific rigor dictates that we discard ideas and practices that are shown to be false-hypotheses, and encourage practices that conform to scientific evidence.

Recommendations

With consideration for the availability of generalizable normative data and generalizable knowledge of test accuracy characteristics, the proposed use of the AFMGQT should be endorsed without the need for additional ethics protocols pertaining to the notification to examinees and referring professionals regarding the use of experimental methods for which normative data and published evidence of validity and effectiveness are not available. It is our conclusion that it does not serve the profession or the community to attempt or maintain a dogmatic position that is inconsistent with scientific evidence regarding the role or need for outside-issue questions. We note, however, that potential confusion exists surrounding the interpretation of test results obtained using the AFMGQT format. Test data obtained using this format are traditionally interpreted with the assumption of independent criterion variance, while the proposed use involves the assumption of test questions for which the criterion variance is assumed to be non-independent. For this reason, we suggest that examiners engage a standard practice of including language in examination reports to clarify testing assumptions regarding independent or non-independent criterion variance. We further suggest that report language should inform referring professionals of the basis of normative data to describe the probabilistic results that underlie categorical conclusions regarding deception or truth-telling. We offer the assistance of members of the APA research committee in the formulation of standard report language and training to achieve these suggested notification and reporting objectives.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, 40(4), 196-305.

Backster, C. (1962). Methods of strengthening our polygraph technique. *Police*, 6, 61-68.

Backster, C. (2001). A response to Krapohl & Ryan's "belated look at symptomatic questions." *Polygraph*, 30, 213-215.

Capps, M. H., Knill, B. L., & Evans, R.K. (1993). Effectiveness of the symptomatic questions. *Polygraph*, 22, 285-298.

Department of Defense (2006). Federal Psychophysiological Detection of Deception Examiner Handbook. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Reprinted in Polygraph, 40(1), 2-66.

Honts, C., Amato, S., & Gordon, A. (2004). Effects of outside issues on the comparison question test. *Journal of General Psychology*, 131(1), 53-74.

Honts, C. R. & Peterson, C.F. (1997). Brief of the Committee of Concerned Social Scientists as Amicus Curiae *United States v Scheffer*. Available from the author.

Kircher, J. C., Horowitz, S. W. & Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, 12, 79-90.

Krapohl, D. J., & Ryan, A.H. (2001). A belated look at symptomatic questions. *Polygraph*, 30, 206-212.

Krapohl, D. J. & Ryan, A.H. (2001). Final comment on the belated look at symptomatic questions. *Polygraph*, 30, 218-219.

Matte, J. A. (2001). Reply to rejoinder by Donald J Krapohl and Andrew H Ryan. *Polygraph*, 30, 220-222.

Matte, J. A. (2001). Comments on Krapohl & Ryan criticism of Capps, Knill & Evans research on symptomatic questions. *Polygraph*, 30, 216-217.

Nelson, R. & Handler, M. (2013). A brief history of scientific reviews of polygraph accuracy research. *APA Magazine*, 46(6), 22-28.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40, 67-78.

Raskin, D. C., & Honts, C. R. (2002). The Comparison Question Test. In M. Kleiner (Ed.), *Handbook of Polygraph Testing*. Academic Press.