February 11, 2009

Dear Colleagues,

I was asked to conduct an "independent peer review" of the attached manuscript entitled "Voice Stress Analyser Instrumentation Evaluation." I chose to proceed as if this effort were a typical panel of referees for a manuscript submitted to the Journal *Science.*

Accordingly, the manuscript was redacted of identifiers of the authors. I selected four referees with the following qualifications:

- A PhD Biosecurity Expert, experienced in "personnel reliability" programs
- A PhD Decision Scientist from RPI
- A PE, MS Biostatistician, specializing in study design
- A PhD Medical Psychologist, experienced in criminal and behavioral psychopathies

I have attached the guidelines for review used to assist each referee in formulating comments. Referees will remain anonymous, and I submit the attached aggregated summary comments as the final peer review. Individual reviewer comments are also attached.

No one in this process was compensated; however one reviewer declared a potential conflict of interest in that he believes he knew the identity of the manuscript authors.

Thank you. You may contact me directly at tom@whalen3.org.

Sincerely,

Thomas Whalen
Professor of Decision Science
J. Mack Robinson College
Georgia State University
Atlanta, GA 30303-3083 USA

Attachments:

Guidelines for Reviewers and Summary Review
Individual Referee Comments
Manuscript sent to reviewers

Summary Review of the draft manuscript "Voice Stress Analyzer Instrumentation Evaluation"

Four referees were asked to review a manuscript as if they were reviewing a paper submitted for publication in a scientific journal. The following specific questions were posed to them:

How thoroughly conceptualized are the criteria for comparing the two methods?
To what extent is prior research taken into account and properly credited?
What prior research, if any, would have materially strengthened or weakened the conclusion if it had been considered?
How sufficient are the data to support the conclusions?
What additional data, if any, ought to have been collected?
Are the methods used to analyze the data appropriate?
What changes, if any, would improve the methodology?
Are the results of the analysis properly interpreted and characterized?
What changes, if any, should have been made to the interpretation and characterization?
To what degree are the conclusions supported by the study?

The reviewers are all in agreement that the manuscript "Voice Stress Analyzer Instrumentation Evaluation" and the work it reports contain a number of serious scientific flaws; as a result, the claims made in the manuscript are not adequately supported by the evidence presented. This work does not merit publication in its current form.

Reviewer 1:

The dependent variable in this study is a binary classification by experts of whether a particular speech sample shows "blocking," which the makers of the devices being assessed say is an indicator of deception and/or stress. A number of 2 by 2 comparisons show the proportion of speech samples classified as high or low blocking versus one treatment designed to induce stress or deception versus another designed not to. When a speech sample from a low stress or low deception treatment is classified as showing blocking, this is reported as a false positive. The principal conclusion drawn by the authors is that the number of false positives is so high that blocking as categorized using these devices is valueless in detecting deception or stress.

The biggest weakness in this conclusion is the fact that all the speech samples came from "carrier" sentences such as "This is a position I am very comfortable with because I have thought about it for a while and it makes sense" embedded in the middle of a prepared written statement. Statements like this were variously embedded in a true or false statement, concerning neutral or highly charged matters, and with or without electric shock. But the sentence itself had an element of falsehood since it always referred to material created by the experimenters and read aloud by the subject as if it were the product of considerable thought by the subject. Thus, the data, in particular what the authors refer to as false positives, do not support their conclusions and in fact may tend toward contradicting them.

The use of the carrier sentence came from an overzealous but misguided desire for objectivity. If the person doing the classifying worked from the graphic alone, it is only necessary to conceal which treatment the sample came from to permit analysis of any desired comparison using speech that carried the actual true or false information. If the person doing the classifying had to hear the speech sample, only comparisons between charged and neutral topics might require the use of a carrier.

It is also noteworthy that in the "jeopardy" deception condition, the strongest motivation of the subjects is for their friends whom they believed would hear the recordings NOT to think they (the subjects) actually believed what they were saying, so the motivation was actually the antithesis of an intent to deceive.

In conclusion, the claims of the paper are not supported.

Reviewer 2:

It appears, based on the emails included at the end of the manuscript,that V (the equipment manufacturer) and the Study Team were unable to agree on the study parameters. In the email of 11/17/05, V expresses concern about whether the voice samples to be used "reflect a true intent to deceive as measured by LVA." Nonetheless, the email continues, V still wants to have the samples analyzed. So at this point, it is not even clear what is being measured, since V and the Study Team seem to be working at cross purposes.

V asserts that in order to properly test LVA, the Study Team must collect voice samples from real-life situations where people have a "true intent to deceive." V points out that a prior study was criticized for "the use of artificial attempts to create an equivalent to real-life deception." The Study Team protests that real-life situations cannot be used, since there is no way to verify the speaker's intent.

Subsequent emails from the Study Team contain requests for V to quantify descriptive phrases such as "higher than normal" and "increase or decrease dramatically."

In Summary: The conclusions are not supported by the study because V and the Study Team never agreed on what should be measured, what was being measured, and how to measure it. Instead, they agreed to disagree and proceeded anyhow, which invalidated the conclusions at the outset.

Recommendation: Conduct a follow-up study, with both V and the Study Team (preferably a different study team, in order to have a fresh start) agreeing beforehand on a test protocol and metrics that will produce results valid for determining the effectiveness of LVA. Both parties need to be satisfied with the protocol and metrics before proceeding any further.

In addition, it would be helpful if V worked on the following issues prior to the follow-up study:

1. More clearly and quantitatively defining the appropriate metrics for evaluating LVA

2. Defining the data that need to be collected to derive the metrics

3. Developing a study protocol that collects the data under circumstances that are conducive to a valid evaluation of LVA (in particular, how a study could be done using data from real-life situations and how "intent to deceive" might be objectively evaluated).

Thank you for giving me the opportunity to review this work.

Reviewer 3:

There is no way that this manuscript, if presented for the very ordinary and normal peer review process for scientific publication, would pass even a perfunctory "smell test." There's a fatally flawed approach to science that is best described as "I wouldn't have seen it if I hadn't believed it." This manuscript reeks of such presumption.

Although it is "noble" that the authors disclosed (p. 20; Appendix C) that their experimental methodology for testing of the LVA system was *never* (emphasis added) in concurrence with technical requirements posed by the product manufacturer, this fact by itself completely negates any of the conclusions purported by the investigators. It is simply impossible to draw meaningful conclusions from study results when the study methodology leaves open serious doubts as to whether or not the intrinsic operational/functional mechanism(s) of the tested technology was(were) compromised or breached.

Furthermore, this problem is exacerbated by the extensive differences described in the samples, calibrations, and other of "several relationships" (p 19) required to compare results of testing with the LVA technology to other tested technologies.

It is rather like trying to report comparative weight measurements in an experimental system in an orbiting spacecraft, utilizing a variety of scales -- one of which cannot be confirmed as being operational in a weightless environment.

Reviewer 4:

After reading this overly-long manuscript, one is left with the impression that (a) there is some interesting technology in play, (b) someone should design an appropriate study to evaluate its performance, and (c) this is not that study.

Most, if not all, of the myriad tables reveal comparable rates of true and false detection of alleged "deception" by the study subjects. Many of the tables indicate differences so minor in comparison to the magnitudes that any difference other than normal statistical variation would seem implausible. Without the need for sophisticated statistics or modeling, it is clear that no significant differences would be found by simple tests such as Chi-squared. Indeed, if more complex tests, such as the repeated-measure analysis mentioned later in the text, were reported to indicate differences, a rational reader might ask: how and why was the study design so confounded that the crude results are so comparable, if indeed there is a difference.

One wonders (a) why are there so many tables and (b) why are the rates so high in many of the tables. On the first point, a more concise report would include some form of summary table plus perhaps all the others relegated to appendixes. The table values lead to a closer inspection of the underlying data, with the question in mind: were the tests set up such that the majority of the statements made by the subjects were, in fact, lies - or did the machine just "think" they were? This question might be answered by the inclusion of counts, not just percentages, in the tables. The closer scrutiny of the data opens a new can of worms: the fundamental study design.

The usual advantage of a designed, experimental study, as opposed to an observational or retrospective study from data intended for other purposes, is the ability to control factors which might plausibly confound the results. In this situation, obvious confounders include: (1) the order of testing - some people telling "lies" earlier or later in the process, (2) the fundamental difference in the technologies, and (3) the fundamental differences in the various test groups. On the first point, order, classical biological testing frequently incorporates a cross-over design for exactly this reason. On all three points and possibly others given the discussion of differences of opinion between the investigators, the principal investigators missed the opportunity to deal with these issues at the design phase. Instead, the work seems to have proceeded with a hope and prayer that such basic issues would be addressed in the analysis. It appears that possibly they were not.

Getting back to the data, there were issues about subjects' reading and machines/operators' interpreting paragraphs with the one technology and "reading" monosyllabic yes/no responses with the other technology. It is not clear how the study design intended to reconcile these differences, if, for example conflicting results had been found, say good lie detection for the yes/no's and not for the paragraphs. An ideal design would include passages, on the one hand, and questions and answers, on the other, that created comparable senses of truthfulness. Direct comparison might then have been possible.

Jeopardy was described operationally but was not explicitly defined, say by degree of variance in the usual physiological tests. Given the apparent (by accepting the data at face value) overall result that the detectors do not detect, and notwithstanding the other data issues mentioned above, the jeopardy becomes suspect. Was there in fact any sense of jeopardy, or were the subjects essentially reading from a script? Was there so much sense of jeopardy that that alone explains all the relatively high detection rates? Looking at the paper as a whole, these seem obvious questions that should be addressed more credibly in the report. Alternatively, these are fundamental flaws in the study, that along with the design and data issues, make it unsalvageable.

The subjective discussion of study bias which appears as an after-thought in the manuscript is totally out of place and worrisome. Those issues should have been long resolved through study design, including pilot studies if necessary for especially "sticky" issues. The rambling discussion detracts from the credibility, rather than adding the "next step" insight which would be conventional at that point in the manuscript.

The conclusions do seem supported by the data if taken at face value. Yet the conceptualization of the study, the description of the data, the experimental methods, and the analysis methodology all lack credibility. The overall conclusions therefore cannot be supported by the study as reported.