"I know what a worm looks like," says Dad, adding a note of pique to the incredulity, "but, you couldn't possibly have a worm in your corn. I bought that corn myself on the way home from work from the finest produce store in town. It was picked just a few hours before I bought it. Not only that, but that happens to be Golden Glow corn which, you may be interested to know, was developed by the U.S. Department of Agriculture. It was also judged to be the best sweet corn by a panel of corn experts from all over the United States. I saw an article in the *Corn Growers Weekly* just the other day that extolled the virtues of Golden Glow corn and it noted specifically that it is 'worm free.' And another thing, *Consumer's Guide* rated it 'number one' out of all the corn tested. And remember, Son, corn is as American as apple pie."

"But Dad, I can see....."

"Don't interrupt me," Dad shot back, "I'm trying to enlighten you about Golden Glow corn so you won't make the same mistake again. Just to show you how misguided you are, we'll conduct a poll right here. Mother, you ate the corn and I didn't hear any gagging sounds coming from your side of the table. In fact, I would have to say that you had a kind of ecstatic expression on your face as you were eating."

Mrs. Fox couldn't admit that she had actually just kicked her shoes off under the table and was rubbing her feet together. Dad plowed (sorry) ahead.

"How would you rate the corn on a scale of 'Excellent,' 'Great,' and 'Above Average?'"

"Well, John," she began tentatively, "I'm certainly no expert on corn, but I guess I would have to say it was 'Above Average.'"

Dad turned to Amy Fox, who was just getting ready to bolt from the room when she heard those unmistakable stentorian tones emanating from Dad's side of the table. It was too late.

"Well, how does Daddy's little trouper rate this wonderful corn?"

"Gee, Dad," she said, as she watched the green thing on Tommy's plate make its way over the edge, "I guess it was OK, I mean, you know, it was really, you know, not all <u>that</u> bad."

"Way to go," Dad exclaimed, "another rave review for Golden Glow! Well, Son, what do you have to say now?"

Tommy looked down at the table in time to see "it" making "its" way toward the butter dish. "Well, I guess I feel like I must have made a mistake or something. I guess sometimes we don't really know what we think we know - I feel kind of, well, stupid, I guess."

"Well, Son," said Dad with a note of family pride in his voice (he <u>was</u> musically inclined), "we all make mistakes. I'm just glad that you now know how to sort out the real from the unreal, the important from the unimportant. I'm proud of you, Son."

Dad left the room with a(nother) note of authority in his stride. Sis held a napkin to her mouth and headed for the bathroom. Tommy looked at his mother with a mixture of "Aw, Mom, how could you" and "I understand" on his young, but rapidly aging, face. Mom carefully squeezed the worm (which was about to enter the stick of butter) into the folds of her napkin and headed for the kitchen garbage disposal.

---

The moral of this tale is one, some, all, or none, of the following:

- Worms are in the eye of the beholder.
- Worms are a metaphysical concept, best left to philosophers.
- Some people know worms when they see them and even, on occasion, have suggestions for getting rid of them. For some reason, this characteristic often makes other people very nervous.
- Once you have bitten into a worm, it is really hard to forget it.
- What you find has almost everything to do with where you look.

---

# Worm in My Corn: Reply #2

Harris Shettel
Rockville, MD

For those readers who may be completely mystified by Reply #1 (I must admit, however, I *did* enjoy writing it!), here is a somewhat more traditional and substantive discussion of things that trouble me about the St. John article. First, a few words about the worm for the more literal minded readers.

The analogy between denying the reality of the worm in the corn and the argument put forth by St. John that we are looking in the wrong place (visitor behavior as it relates to exhibits) to learn about the reality of exhibit effectiveness/ineffectiveness, seemed to me to be quite apt. I thought of other analogies that would have served as well – not using patients to find out about a new drug; not using customers to find out how successful a new product will be; not using trainees to find out if a training course is achieving its objectives; not using students to find out how well a new curriculum is teaching – but since St. James used corn to help demonstrate the "folly" of using experimental procedures to learn about exhibits, I thought the "non-existent" worm in the corn was the analogy of choice.

On to other matters. To point out, as St. John does, that in using the experimental research paradigm one must make a distinction between statistical significance and practical significance is to state the obvious. But, to then conclude that the experimental approach is almost by definition suspect ("limited at best and wrong-minded at worst"), is to take a very large and unfounded leap.

Sir Ronald A. Fisher, a recognized pioneer in the development of statistical tests of significance, was among those who acknowledged the agricultural heritage of such testing (Mark St. John being another one!). Fisher noted in his small but influential book, *Statistical Methods and Scientific Inference* (1956), in a chapter called "Some Misapprehensions About Tests of Significance," that such tests are often used, among other things, to "recognize... the reality of the response to manurial treatment of a cultivated crop" (p. 76). These tests, he went on, "are constantly in use to distinguish real effects of importance to a research programme from such apparent effects as might have appeared in consequence of errors of random sampling, or of uncontrolled variabilility of any sort..." (p. 76).

Of course, since 1956, much more sophisticated versions of such statistical tests have been developed, allowing one to measure very complex and multi-dimensional interactions, and they are now an indispensible part of the research programs in all of the physical, biological, and social sciences. No one, to my knowledge, has found a more reliable and valid way to build and test theories, models, and hypotheses, and that applies to theories, models, and hypotheses about exhibits!

But, Fisher, among others, also clearly saw the danger of indiscriminately tranferring parametric statistical procedures to areas that do not lend themselves as easily to meeting the required assumptions, and specifically mentioned judging the effectiveness of military equipment purchased for the Royal Navy as an example. He expressed his concern as follows:

> "...a considerable body of doctrine has attempted to explain, or rather to reinterpret, these (significance) tests on the basis of quite a different model, namely as a means to making decisions in an *acceptance procedure*. The difference between these two situations seem to the author many and wide" (p. 77). [Italics added.]

He then went on to point out how the practical needs of the real world do not necessarily lend themselves to the rigors of a long series of controlled research studies. In fact, he said, it is the "objective empirical reality" of things that actually exist that drives these acceptance procedures, not a series of mathematical assumptions such as those upon which parametric statistical inferences are based. In short, when you have a worm in your corn, you do not need tests of significance to prove it!

What we are talking about here, in effect, is the distinction between research and evaluation, a distinction that St. John fails to make (although his title says "evaluation," his opening discussion is about "research"). I think that this is precisely what Fisher was talking about, although he did not use the term "evaluation" ("acceptance procedures" sounds very close!). The model for most of the exhibit evaluation studies that have been carried out is not the treatment vs. non-treatment experimental paradigm using parametric tests of significance that St. John refers to and (in true strawman fashion) ridicules, but the concept of *criterion-referenced measurement* that was introduced in the early 60s by Glaser and Klaus (1962), and since then has evolved to be the accepted model of assessment/evaluation in many areas, including military/industrial training and (very reluctantly) education. It is, essentially, the same as the goal-referenced model that many of us in the exhibit evaluation field have been talking about and using since the mid-60s.

The central feature of the criterion-referenced measurement paradigm is the *rationale* of the thing being evaluated. All exhibits have a rationale, but most of them are never stated explicitly. A somewhat simplistic rationale that probably applies to most, if not all, exhibits and public programs could be stated something like this: "Our institution has expertise in subject X (and perhaps objects). We believe that visitors to our museum would be interested in/gain something from/benefit from, an exhibit on the subject of X. We have the funds to do an exhibit on X, so let's put it in our planning cycle (or ask for government funding) and do it."

From an evaluation point of view (as well as from a planning point of view) this rationale is extremely weak and could not serve as the basis for designing a study. A more explicit and useful rationale would define in detail the notions of "visitor," "interest/gain/benefit," and "X." The audience needs to be identified in terms of age, education, background, prior knowledge, etc., and "interest/gain/benefit" are turned into behavioral, cognitive and/or affective outcomes for the defined audience (or audiences). Exhibit content, story line, and design features are then planned on the basis of these decisions and also on data obtained from front-end studies designed to get a better handle on what the audience(s) are *really* interested in, as well as what they already know, and, more importantly, what their misconceptions are. Evaluation (both formative and summative/remedial) is carried out against the criteria that are thus established to see if the assumptions made were in fact correct, nearly correct, or not correct. Problem areas thus identified are dealt with by making adjustments to the exhibit and then retesting it.

Various statistical procedures are, of course, often used to determine the "goodness of fit" between the established criteria and the actual results obtained from a representative sample of visitors from the target population. Statistical procedures may also be used to determine if an "improvement" made in the exhibit is a "real" one, real being defined as a practical difference and not just a statistical difference. The kinds of measures used are often of the non-parametric type (such as the chi-square test) where assumptions of normality in the data are not required. Parametric measures (such as the t-test or the analysis of variance) are also used in evaluation work for their increased power and flexibility. Very often, however, simple descriptive statistics are quite adequate to show that there is an unacceptably large gap between what the aims of the exhibit were and what the target visitors who attended to the exhibit were "getting out" of it, or to show that a change made resulted in an improvement in visitor responses or behaviors.

Note that the rationale of an exhibit also includes *attracting* and *holding* visitors, not just communicating its message to them. This is the "interest" factor that was noted earlier. For this reason, evaluation studies almost always include data on attracting and holding power. You certainly cannot learn anything from even the most carefully prepared exhibit if you do not go to see it and attend to it. (On the other hand, I have evaluated exhibits that had close to 100% attracting and holding power that actually communicated the *opposite* of the intended messages!)

These types of concerns open up a whole range of issues that have to do with why people come to museums and (even more importantly) why they don't, how they interact socially while they are in the museum, not to mention their previous experiences with museums and their image of what a museum is. These issues are still not well enough understood, but they are being addressed by more and more of those working in the visitor studies field. (St. John's discussion of anthropology/ethnography is relevant to this area and it is one that more and more visitor study practitioners are investigating.)

But, I continue to maintain that exhibits are the core of the museum (educational) experience. Thus, an exhibit that does not attract, hold, or communicate to its audience as it was intended to do, is by (my) definition, an exhibit with a problem. And, contrary to St. John's assertion that our evaluation studies do not tell us *why* something happens, only *that* it happens, the information obtained from visitors very often points with exquisite clarity to the nature of the problem. If the essential message of an exhibit is in a text panel (as it so often is) and very few visitors read that panel, it is not difficult to conclude that the problem has something to do with that panel. If we ask some visitors to read the panel and then ask them to tell us its basic message, we have further isolated the nature of the problem as between "We need to put the panel where people can see it" and/or "We need to rewrite the text so that people can understand it." I think that our evaluation studies are, in fact, building a good base of empirical knowledge, that, along with the experimental studies that have been done (and, hopefully, more and better ones *will* be done), are providing us with some very useful heuristics and even the beginnings of a theory of exhibit/visitor/museum interaction.

Mistakes have certainly been made. I am sure that trivial statistical findings have no doubt been touted as meaningfully and practically significant. Ad hoc studies carried out for single exhibits (or even "pieces" of exhibits) have, on occasion, been used inappropriately as a basis for making sweeping generalizations. Evaluators have even been accused of stifling creativity by creating a climate in which it is better to be "safe" than take a chance on something new or innovative that may not "work." (I have heard this argument for over twenty years and I still don't understand why meeting certain agreed-to educational goals and being creative are mutually incompatible. Architects and bridge builders, for example, have to meet very strict criteria related to structural soundness, etc., but that does not seem to inhibit them from continually coming up with unique and fresh designs. Even that hotbed of creativity, advertising, cannot ignore the "bottom line.")

In short, I simply cannot accept the notion that we are fundamentally flawed in the work that has been done under the general headings of evaluation and research, nor do I think that we have anything of real value to gain from substituting any of the metaphors suggested in St. John's article. I cannot forget the results of the first exhibit study I did in which exhibit "experts" of various kinds consistently disagreed with each other about what constitutes an effective exhibit (Shettel, 1968). For them, "worms" were clearly in the eye of the beholder. Since that study was done, we have made real progress by bringing the visitor into the research and evaluation process. Mark St. John notwithstanding, I see no reason to take the visitor out!

## References

Fisher, Sir Ronald A. (1956). *Statistical methods and scientific inference.* New York: Hafner Publishing Co.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, pp. 421-427.

Shettel, H. (1968). An evaluation of existing criteria for judging the quality of science exhibits. *Curator*, 11(2), 137-153.