# Improving Methodology in Spreadsheet Error Research

Raymond R. Panko
Shidler College of Business
University of Hawai`i
2404 Maile Way
Honolulu, HI 96821
001.808.377.1149
Ray@Panko.com

## ABSTRACT
Too much spreadsheet research is unpublishable in high-quality journals due to poor methodology. This is especially a problem for computer science researchers, who often are untrained in behavioral research methodology. This position paper reflects the author's experiences in reviewing submissions to information systems and computer science journals.[1]

## Categories and Subject Descriptors
K.8.1: Spreadsheets. D.2.5 Testing and Debugging.

## General Terms
Experimentation, Verification.

## Keywords
Methodology. Spreadsheet Experiments, Experiments, Inspection. Sampling, Statistics

## 1. INTRODUCTION

For a number of years, computer science journal editors have taken to sending me articles to review that involve experimental and other methodology. It is frustrating to review these studies because they often show a weak understanding of methodology. Fatal methodological errors are too common, and errors that hobble the use of results are even more frequent. In spreadsheet error research, methodological issues have been particularly common in papers by computer scientists. Based on my experience, this paper presents some prescriptions for improving spreadsheet error research. We will look at issues in inspections (audits) of operational spreadsheets, spreadsheet development experiments, and spreadsheet inspection experiments.

## 2. INSPECTIONS (AUDITS) OF OPERATIONAL SPREADSHEETS

Several studies have inspected corpuses of operational spreadsheets to look for errors. Many studies call this auditing, but auditing is a sample-driven statistical analysis method for developing an option about quality in development. Audits are not comprehensive error detection tools.

### 2.1 Respect Human Error Research

Inspection methodologies often fail to reflect the fact that software and spreadsheet error rates are similar. Consequently, spreadsheet methodologies tend to ignore the rather vast literature on code inspection. By code inspection standards, most spreadsheet inspection methodologies do look like mere audits. They lack the required initial understanding of the spreadsheet, are undertaken on whole spreadsheets instead of modules, use single inspectors, and so forth.

### 2.2 Don't Trust. Verify.

Spreadsheet inspection methodologies are rarely verified. Instead, they tend to be refined until the researchers "feel good" about them. To verify the effectiveness of a methodology, it is important to have multiple inspectors independently use the same methodology to inspect the same spreadsheets. Comparing errors from multiple inspectors can indicate relative effectiveness in finding different types of errors. If the methodology is strong, cross-analysis can even give an estimate of errors remaining.

### 2.3 Report Time Spent

Time spent in testing is important in assessing human error research. It is important to reveal inspection rates for individual spreadsheets—both time in total and time as a percentage of size expressed in multiple ways, such as all cells, all formula cells, unique formulas, and so forth. If a spreadsheet inspection method has multiple phases, time in each phase should be reported.

### 2.4 Understanding the Spreadsheet First

Spreadsheets are not self-documenting. It is important for inspectors to be given a thorough explanation of the spreadsheet's detailed logic before they begin testing.

## 2.5 Report Error Seriousness

The seriousness of errors—at least the most serious error found—should be assessed. Seriousness should be reported by size of each error on monetary or other scales, percentage size of the error relative to the size of the correct value, seriousness of the error in its context, and risk created for the organization. Context must be understood well. In annual budgeting, small errors can be very damaging, while in major one-off projects such as the purchasing of another company, errors would have to be large compared to the results variance caused by uncertainties in input numbers.

# 3. DEVELOPMENT EXPERIMENTS

In development experiments, participants create spreadsheet models based on requirements in a word problem. To date, we have done well in estimating cell error rate ranges during development. However, there is much more we need to do.

## 3.1 Use New Tasks

Spreadsheet development experiments have only used a few tasks. We need to do development experiments with more tasks to be confident about typical cell error rates. The widely used Wall and Galumpke tasks have different error patterns. We need to try new tasks to see if new patterns emerge. The Wall task is especially problematic because it was designed to be extremely simple and almost free of domain knowledge requirements. Participants make very few errors on the Wall task.

## 3.2 Have Adequate Task Length

Errors are rare in spreadsheet development. Tasks need to be relatively long or there will be too few errors to analyze. One way to address this is to have subjects do multiple tasks in a balanced design and to analyze errors in the total multitask sample.

## 3.3 Go Beyond Student Samples

We also need to do studies on people with different levels of experience in spreadsheet development to ensure that spreadsheet research does not suffer from being the science of sophomores.

## 3.4 Test Prescriptions for Safety and Effectiveness

We need to move beyond simply claiming that certain prescriptions (such as have a separate assumptions section) and certain tools are good ideas. We must test them to see if they really are "safe and effective." We cannot just build tools and make claims about why they will save the world. Prove it.

## 3.5 Go All the Way to Error Reduction

Showing that users like it or showing that a tool can help point to earlier cells is not enough. Does it reduce errors? If not, who cares?

## 3.6 Use Ample Sample Sizes

Sample sizes must be large—at least around 30 to 50 participants per condition. Otherwise, statistical analysis is unreliable. The minimum number should be determined empirically, by a power test.

## 3.7 Avoid Friends and Family Samples

We also need clean samples. Mixing highly experienced professionals with rank novices in the sample requires far larger samples for statistical validity.

## 3.8 Do Rigorous Random Assignment to Conditions

Doing rigorous random assignment to the control and treatment groups is mandatory and critical. This must be done on the basis of individuals. We cannot assign whole class sections to different treatments. Nor can we place earlier arrivers in one condition and later arrivers in another condition.

## 3.9 Use Nonparametric Statistics

It is important to use nonparametric statistics because errors do not follow the normal distribution even roughly. Transforming data so that they are pseudonormal and then applying traditional parametric statistics is not acceptable today.

## 3.10 Be Generous in Presenting Statistical Results

When giving results, do not just give bare minimum result numbers like means, medians, and standard deviations. Show the full results matrix generated by statistical analysis programs. Also, in comparisons, give overall numerical differences. Do not just say that a difference was statistically significant without giving the numerical differences or correlations.

# 4. INSPECTION EXPERIMENTS

Inspection experiments should follow the advice in both previous sections. It is wise to avoid seeded errors and go with data from actual development experiments. (The author has such a corpus.)

## 4.1 Higher Error Rates

One good thing is that human error detection rates are worse than error commission rates, so sample can be a little smaller and still generate enough errors. However, statistical analysis is misleading with less than about 30 subjects per group and rigorous subject randomization.

## 4.2 Test for Safety and Effectiveness

Again, we need to go beyond simply measuring error detection rates and move to testing alternative methods for finding errors. If we test only two methods—such as doing nothing and using a particular method, then we double the required sample size and must be extremely careful about random treatment assignment. Effects size is also critical in selecting sample sizes.

# 5. CONCLUSION

We need to stop touting untested prescriptions and tools if we are to put our field on a scientific footing. We must scrutinize prescriptions for safety and effectiveness, and we must do so with exemplary methodology. We also should be balanced in our presentation of results. Everything has strengths and weaknesses. Our results should be honest about weaknesses. Obscuring methodology is a professional sin.