Chapter 11
Errors in Operational Spreadsheets

Stephen G. Powell
Dartmouth College, USA

Kenneth R. Baker
Dartmouth College, USA

Barry Lawson
Dartmouth College, USA

ABSTRACT
Spreadsheets are used in almost all businesses, for applications ranging from the mundane to the mission-critical. Errors in the data, formulas, or manipulation of spreadsheets could be costly, even devastating. The received wisdom is that about 5% of all formulas in spreadsheets contain errors, and this rate is consistent across spreadsheets. However, this estimate is based on five studies, some of which are quite informal, and a total of only 43 spreadsheets. Our research was designed to deepen our understanding of spreadsheet errors. Specifically, we address three questions about errors in operational spreadsheets: what is the average cell error rate, how does it differ among spreadsheets, and what types of errors are most prevalent? We created a spreadsheet auditing protocol and applied it to 50 diverse operational spreadsheets. We found errors in 0.9% to 1.8% of all formula cells, depending on how errors are defined. We also found that the error rate differed widely from spreadsheet to spreadsheet.

INTRODUCTION
Errors in software have been a problem since the beginning of the computer era. The discipline of software engineering (Sommerville, 2004) was created in part to develop methods to minimize errors in software code. Professionally developed software is now generally created and debugged using disciplined approaches from software engineering.

The advent of the personal computer in the 1980s increased by orders of magnitude the number of people who worked closely with computers. Just as professional programmers were learning to use disciplined approaches to programming, millions of so-called end-user programmers be-
gan to create their own programs in the form of spreadsheets. One of the appealing features of spreadsheet programming on a personal computer is that the end user can work independently of the constraints of a professional programmer or an MIS group. This allows the end user much more freedom and speed in completing the business task at hand. But very few end users have any training for the programming tasks they attempt, and many have little appreciation for the risks inherent in a casual approach to programming.

As spreadsheets have spread throughout business, they have increasingly been used for mission-critical applications (Croll, 2005; Grossman, Mehrotra, & Özlük, 2005). At the same time, evidence has accumulated that many spreadsheets have errors (Panko, 2005) and that sometimes these errors cost the organizations that use them millions of dollars (EUSPRIG, n.d.). End users and organizations that rely on spreadsheets generally do not fully recognize the risks of spreadsheet errors. In fact, spreadsheets tend to be somewhat invisible, both as corporate assets and as sources of risk.

Although the received wisdom holds that errors are prevalent in spreadsheets, our review of the literature suggests that this conclusion is based more on casual empiricism than solid research. In fact, researchers have produced only limited information on the types of errors that occur, how frequently they occur on average, and how the error rate varies from one spreadsheet to another.

The current study was designed to improve our understanding of spreadsheet errors by testing a large sample of spreadsheets in actual use by organizations. We developed a detailed auditing protocol and trained a group of researchers to apply it consistently. Using an explicit protocol is important for two reasons: (1) it allows other researchers to replicate and improve on our work; and (2) it contributes to the development of improved auditing procedures, which is important in its own right. The auditing procedure is described in detail in Powell, Baker, and Lawson (2008b).

We begin this article with a summary of previous work on spreadsheet errors which comes from field audits and laboratory experiments. Then we describe the design of our study and the sample of spreadsheets we audited. Our results are summarized in terms of error instances (the occurrence of a single type of error) and error cells (the number of cells affected by a single error instance). We also categorize the errors as to whether they generate wrong numerical results. We report the total number of errors we identified, as well as how they were distributed across error types and across spreadsheets. This is the first data on errors and their distribution to appear in the research literature that can be replicated with parallel studies using similar methods.

PREVIOUS WORK ON SPREADSHEET ERRORS

Since our interest in this article is primarily with the classification and frequency of errors, we will discuss only those questions here. In a related article (Powell, et al., 2008a), we provide a review of the literature on all aspects of spreadsheet errors.

Types of Errors

Before we can productively study spreadsheet errors, we need to define and classify the types of errors we seek. Classification is useful because it helps us to understand what is common among errors. It also allows us to compare different studies of errors, and eventually to understand the causes of errors and to develop cures.

At the outset, we make a distinction between errors in the use of a spreadsheet and errors in the cells of the spreadsheet itself. A spreadsheet can be perfectly correct but used to solve the wrong problem, or its results could be misinterpreted. More concretely, a correct spreadsheet can lead to an error if sorting is done improperly, or if formulas are overwritten, or if incorrect data were entered.