RANGE NAMES
A Risky Practice in Spreadsheet Development?

Ruth McKeever and Kevin McDaid
Dundalk Institute of Technology, Dundalk, Co., Louth, Ireland

Keywords: Spreadsheets.

Abstract: The use of range names to improve spreadsheet development is advocated by both academics and practitioners, however there is a clear absence of supporting scientific evidence. This paper describes the latest in a series of experiments that examine the impact of range name structures on spreadsheet reliability, and formula development time. The aim of this paper is to compare the reliability of simple spreadsheet formulas developed by intermediate users through both cell references and range names. The results are consistent with the findings of previous experiments that, contrary to widespread opinion, the use of range names does not improve the quality of spreadsheets.

1 INTRODUCTION

The importance of spreadsheets cannot be overstated, as there are now believed to be 400 million Excel users worldwide. The importance, and uncontrolled use, of spreadsheets in the financial sector is investigated by Croll (2005). It is reflected in the following quote made in relation to the financial sector in the City of London: "Excel is utterly pervasive. Nothing large (good or bad) happens without it passing at some time through Excel".

Many authors acknowledge that spreadsheet development is programming. For example, Burnett et al (2004) declares that "spreadsheet languages are the most widely used end-user programming languages to date—in fact, they may be the most widely used of *all* programming languages".

Spreadsheets are both powerful and flexible and, as a result, are an indispensable tool in modern business. Flexibility, however, comes at a cost. Any user can become a developer, without any knowledge of risk or testing procedures. This leads to widespread uncontrolled use of poorly developed spreadsheet models. Considerable research has recognised that spreadsheets are rarely developed by professional programmers. For example, Purser and Chadwick (2006) found that 85% of survey participants developed the spreadsheets that they use. This is not to say that the developers are not professionals, but that their expertise lies within their domain rather than in programming. Consequently, spreadsheet development is known to be highly unreliable, and spreadsheets have been linked with many recent high profile and costly errors. Powell et al. (2007) found errors in 94% of spreadsheets, and 1-2% of cells. The reliability of a spreadsheet is essentially the accuracy of the data that it produces, and is measured by the number and magnitude of errors found in the spreadsheet. To improve reliability many expert practitioners advocate the use of range names in the development of spreadsheets. However, there is no empirical research to support these claims.

This study is part of the first empirical investigation into the impact of range names on spreadsheet reliability. It summarises the results of earlier experiments on the impact of range names on spreadsheet debugging and it presents in detail the results of a recent experiment investigating the use of range names in the development of simple spreadsheet formulas. This work is important in that, contrary to published opinion, our evidence suggests that range names have a negative impact on the reliability of simple spreadsheet formulas, and on the time it takes to develop them.

1.1 Spreadsheets

Errors in spreadsheets that lead to bad decisions are often reported in the media, and a list of examples

McKeever R. and McDaid K..
RANGE NAMES - A Risky Practice in Spreadsheet Development?.
DOI: 10.5220/0003510202690274
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 269-274
ISBN: 978-989-8425-76-8

can be found on the European Spreadsheet Risks Interest Group (EuSpRIG) website.

Spreadsheet engineering is frequently described as end-user programming. A 2005 study (Scaffidi et al., 2005) estimates that by 2012 there will be over 13 million end-user developers in the US, compared with 3 million professional programmers. It is acknowledged that end-user developed systems introduce risk into an organization, and these risks can have many influencing factors, for instance, spreadsheet developers do not follow a structured process, many are untrained in the use of processes as in software development, and are rarely aware of the unreliability of spreadsheets. One study found that only 6% of development time is spent testing spreadsheets (Baker et al., 2006).

Spreadsheet researchers all converge on the same findings – spreadsheet use is ubiquitous, and spreadsheet quality is not considered paramount within organisations. Testing is not considered crucial, and critical decisions are made based on unregulated spreadsheets.

1.2 Naming

Many researchers have examined the importance of naming in programming. Keller (1990) found that people who read programs that followed a defined naming scheme found them easier to read, but could not pinpoint why. Jones (2008) raises the issue of incorrect spelling: "if people make spelling mistakes for words whose correct spelling they have seen countless times, it is certain that developers will make mistakes, based on the same reasons, when typing a character sequence they believe to be the spelling of an identifier".

Range names are a feature in Excel that allows a developer to assign a meaningful name to a cell or group of cells. This name can then be used throughout the spreadsheet instead of the cell reference. Names can be easily created, modified, and deleted through the Name Manager, located in the formulas tab of the Excel ribbon.

Practitioners often suggest that range names can make spreadsheets easier to understand and to develop, in books, academic papers, journals, and on websites, illustrated by the following examples:

- "Range names improve reliability. If you need to change references to the range, you only have to change the definition of the range name. Then every formula that uses it will refer to the new address." (O'Beirne, 2005)
- "Clearly, using the Defined Names makes the formula much easier to understand and maintain." (Pearson, 2009)

In contrast, some experts caution against using range names. Panko and Ordway (2005) warn that range names "should be considered potentially dangerous until research on using range names is done." Blood (2006) states that names are unnecessary if the model is well designed, and that range names make it more difficult to audit formulas, as important information becomes hidden. He also criticizes range names for making formulas unnecessarily long.

1.3 Summary

Many researchers have explored spreadsheet errors, in terms of their frequency and causes. It is widely recognised that spreadsheets are unreliable, and range name use is often mentioned as a practice that improves spreadsheet quality. The majority of practitioners are in favour or range names, yet a few vocal opponents remind us that there is no scientific evidence to support these recommendations.

This paper outlines earlier aspects of our research programme, and then describes in detail an experiment to establish the effect of range names on the reliability of basic spreadsheet formulas.

2 BACKGROUND

This study is part of a wider research project that investigates the impact of range names on spreadsheet reliability, in order to assess the feasibility of recommending range names for use in spreadsheets, and is guided by the following objectives:

Objective 1: Investigate the impact of range names on the ability of novice users to successfully identify and correct errors in a spreadsheet.

Objective 2: Investigate the reliability of spreadsheets developed using range names.

Objective 1 was addressed by experiments detailed in Section 2.1. The study presented in this paper begins to address Objective 2 by examining the reliability of formulas developed using range names, as compared to formulas developed using cell references.

2.1 Debugging Experiments

The experiments that addressed Objective 1 were adapted from a design first used in a study by Howe and Simkin (2006), and later used by Bishop and

McDaid (2007). Participants were given a spreadsheet seeded with errors, and were asked to correct any mistakes they could find, directly in the spreadsheet. They were not told how many errors were in the spreadsheet, or what types of errors were included. Their cell clicks were recorded by T-CAT, a "time-stamped cell activity tracking tool".

2.1.1 Experiment 1

Initially a small exploratory experiment was carried out on 21 computing students (McKeever et al., 2009). The spreadsheet, seeded with 42 errors, was first modified to include range names in formulas.

Table 1: Results of exploratory study.

Error Type	No. of Seeded Errors	% Corrected by Participants
Clerical	4	11%
Rule Violation	4	63%
Data Entry	8	64%
Formula	26	47%

When these results were compared with the results of an identical study by Bishop and McDaid (2007) which did not involve range names, it was found that the students in the exploratory trial found 19% fewer formula errors than the students who took part in the trial without range names. Further analysis showed that the exploratory group found 29% fewer cell reference errors, and 24% fewer range reference errors.

The significant difference in results indicated that the inclusion of range names in a spreadsheet does not make it easier to debug. Importantly it prompted a second, better-controlled experiment with more focussed and detailed research questions.

2.1.2 Experiment 2

Based on the results and feedback from the first experiment, a new set of research questions were developed for a controlled second experiment which examined, in a more rigorous way, the impact of range names on the spreadsheet debugging process (McKeever and McDaid, 2010). This experiment was designed to investigate whether users are better at finding errors in formulas when the formulas contain range names as opposed to cell references, and to provide information on the following four cases of error involving range names:

RQ1: The error is due to the wrong range being assigned to a name.

RQ2: The error is due the wrong range name being used in a formula.

RQ3: The formula contains a name, but the error is not due to the name.

RQ4: There are no names in the formula, but names in the spreadsheet.

The spreadsheet used for the first trial was modified for this experiment, to reflect feedback received. The number of names in the spreadsheet was reduced from 152 to 12. The number of errors was reduced from 42 to 39. 29 students took part in this trial, and this time they were divided randomly into control and treatment groups.

The treatment group found 4% fewer errors than the control group, for the errors relating to RQ1 and RQ2; these results are not statistically significant. The treatment group also found 20% fewer errors relating to RQ3, and 12% fewer errors relating to RQ4; these results are statistically significant.

2.1.3 Summary

These findings demonstrate how range names make debugging a spreadsheet more difficult for novice users. Development however, is arguably more important than debugging; if range names can help developers avoid errors initially then less debugging is necessary. For this reason, we now focus our research on the role of range names in formula development.

3 RESEARCH QUESTIONS

To address Objective 2, to investigate the reliability of spreadsheets developed using named ranges, the authors first looked at the reliability of basic formulae developed using range names. The work was guided by the following two research questions:

RQ1: Do users make more mistakes using range names or cell references, when asked to develop a simple spreadsheet formula?

RQ2: Does the time it takes users to develop a simple spreadsheet formula differ for formulas using range names than for formulas using cell references?

Range names, as with programming variables, can be chosen according to various conventions. The work will examine each of the research questions above for each of the following six range naming structures:

a. Where no two names begin with the same word.
b. Where several different names begin with the same word, but end in a different word.
c. Where several different names begin and end in the same words, with a change in the number in the middle of the name.
d. Where names begin with the same word with a

change in the trailing number.

e. Where several different names begin and end with the same word, with a minor change in the word in the middle.
f. Where names do not follow any naming convention, and are inconsistent.

4 METHODOLOGY

To address these questions we decided to isolate a basic formula task, one that the participants would have used many times. It was decided to ask the students to add the values from a number of cells together. The participants would not require domain knowledge to complete the tasks.

4.1 Task Design

The spreadsheet designed for this experiment held six worksheets, each of which contained two tasks. For each task they were asked to calculate the total of a number of specified cells, one task using cell references, the other using range names. This resulted in a total of twelve tasks, six for cell references and six for range names.

The range names in each sheet were developed using a different naming convention, in order to examine each structure defined in the research questions. Examples are shown in Table 2:

Table 2: Naming Conventions.

Sheet	Name Example
Sheet 1	ArnottsSales, ClearysSales
Sheet 2	TopshopGP, TopshopNP.
Sheet 3	HMV2008Profits, HMV2009Profits
Sheet 4	PrimarkTax2006, PrimarkTax2007
Sheet 5	GAPSecWages, GAPSupWages
Sheet 6	PPatchVar, Fixed_Costs_Pumpkin_Patch

Sheet level names were used for this experiment, so that the subject would only be able to view the names relevant to the task on which they were working. There were 264 names in the workbook, and this allowed us to examine the naming structures in isolation without confusing the participants.

4.2 Conducting the Experiment

The participants in this trial were 15 postgraduate students from the Higher Diploma in Computing class in Dundalk Institute of Technology, most of whom had spend a period of time a period in the workplace before returning to education. The participants were considered intermediate users, based on a background survey carried out on a sample of 14 members of the class.

Within subject design was chosen for this study, due to the low number of available participants. Each participant developed twelve basic formulas, six using cell references and six using range names. To avoid any carryover effect that might occur, the order in which the participants carried out the tasks was alternated, by dividing them into two groups. Group A used range names for the first task on each worksheet and cell references for the second; group B used cell references for the first and range names for the second.

After randomly dividing the subjects into groups the researcher explained how to complete the experiment. T-CAT was used to record the time each task took, and the participants were observed throughout.

4.3 Example

The following task was given to the participants in Group A, on Sheet 5.

Task 9: With a formula that uses range names, in cell C28 calculate the total wages for the following employees :
TJ Hughes supervisor
Mackays secretary
GAP secretary
Clintons Card Shops supervisor
Eisengger supervisor
Eisengger secretary
Oasis secretary

One participant made the following erroneous answer to this task:

=TJHughesSupWages+MackaysSecWages+ClintonsCardShopsSupWages+EisenggerSupWages+EisenggerSecWages+OasisSecWages

The correct answer would also include "GAPSecWages".

5 RESULTS

Out of the fifteen participants, only seven completed all tasks correctly, with eight students making at least one error. The following results are based on the individual errors made, rather than the participants who made the error.

5.1 Errors Made

Table 2 shows the number of errors made in each of the sheets on the spreadsheet, according to whether the task included range names or cell references.

Table 3: Experiment results.

Errors	Named Ranges	Cell References
Sheet 1	0	1
Sheet 2	4	0
Sheet 3	3	0
Sheet 4	0	1
Sheet 5	3	2
Sheet 6	2	0
Total	12	4

We expected to find two error types – selection errors, where the wrong range is used in the formula, and omission errors, where a reference is left out. Five of the range name errors were selection errors, and six were omission errors; one cell reference error was a selection error, and three were omission errors. In addition to these errors, another type occurred when participants used range names: in two cases a subject added an extra name to the formula. During task 3 one subject added the name ZaraNP twice, and during task 6 a participant included Costcutter2008Profits and Costcutter2009Profits.

Statistical analysis was conducted on these results, using McNemar's test for paired proportions. This indicates that in two cases, the second and third naming conventions, the subjects were less effective at developing formulas using range names, than with cell references.

5.2 Time Results

Table 3 shows the average time in minutes it took the participants to complete the tasks on each sheet, according to whether the task included range names or cell references. It took on average 0.7 minutes longer to complete the tasks using range names.

Table 4: Average times (minutes).

Errors	Named Ranges	Cell References
Sheet 1	1.98	1.33
Sheet 2	1.95	1.3
Sheet 3	1.45	1.1
Sheet 4	1.35	1.04
Sheet 5	1.31	1.19
Sheet 6	3.19	1.04
Total	1.87	1.17

The times recorded by T-CAT included the assimilation time for both beginning the experiment, and for each sheet. This means that the time it took each participant to complete the second task on each sheet should have been naturally less than the first. Such bias was eliminated through the structure of the experiment and randomisation of participants.

5.3 Findings

Statistical analysis of the significance of the difference in the times to perform each task was performed based on a paired T- test, with a 5% level of significance. These statistical tests support the following statements:

1. Intermediate users make fewer mistakes when developing formulae using cell references than using range names where:
   a. Several different names begin with the same word, but end in a different word.
   b. Several different names begin and end in the same words, with a change in the number in the middle of the name.
2. Intermediate users take less time to develop formulae using cell references than using range names where:
   a. Several different names begin with the same word, but end in a different word.
   b. Several different names begin and end in the same words, with a change in the number in the middle of the name.
   c. Names begin with the same word with a change in the trailing number.
   d. Names do not follow any naming convention, and are inconsistent.

5.4 Issues

One problem with the task was the if the user clicked on a cell that was named while writing a formula, the name of the cell appeared in the formula rather than the cell reference. Unfortunately this is not something that can be caught by the T-CAT macro. The participants were observed throughout the task however, and the researcher is satisfied that they followed the instructions exactly. Any future repetitions of this experiment on larger groups will have to take this into consideration.

Students were used in these experiments. Although this approach can be controversial, studies have shown that students have similar abilities to professionals. As stated previously, it is frequently the case that spreadsheets are, in fact, developed by novice or intermediate users.

The participants in this trial were taught how to use range names, which is not the case for real-world users.

6 CONCLUSIONS

This experiment shows that there is no evidence to support the theory that range names reduce the quantity of errors found in spreadsheets, or make them easier to use. The number of errors made when the participants used range names was higher overall than when the same subjects used cell references. The average time it took to complete each task was also higher when the participants used names.

Importantly, the increase in selection errors illustrates that range names do not help the user to avoid referring to the wrong cell, as is often claimed. The increase in the time it takes to develop a formula dispels the notion that range names make formulas easier to create.

Methods that appear to work for a small number of expert developers must not be presumed to work for other less experienced users who do not have the same experience of errors. This work concludes that, while spreadsheet quality is a real and important issue, the use of range names is not the solution for novice or intermediate level users.

6.1 Future Work

First we plan to repeat this experiment on a larger group of subjects to improve validity. Range names could possibly save developers time when cells are located on different worksheets; this is something we intend to investigate in the future.

More generally, we anticipate the work will look at impact of range names on the entire spreadsheet development process. This is part of a larger study and the there are also plans to extend this experiment to focus on more general spreadsheet development. It is important to examine the performance of expert users with regard to range names.

REFERENCES

Baker, K. R., Foster-Johnson, L., Lawson, B. & Powell, S. G. 2006. *A Survey of MBA Spreadsheet Users*.

Bishop, B. & McDaid, D. K. 2007. An Empirical Study of End-User Behaviour in *Spreadsheet Debugging. PPIG.* Salford 2007.

Blood, A. T. 2006. *Elements of Good Spreadsheet Design: Spreadsheet Design Notes* [Online]. Available: http://www.xl-logic.com/modules.php?name=Content&pa=showpage&pid=2 [Accessed May 2009].

Burnett, M., Cook, C. & Rothermel, G. 2004. End-User Software Engineering. *Communications of the ACM,* 47**,** 53-58.

Croll, G. J. 2005. The Importance and Criticality of Spreadsheets in the City of London. *EuSpRIG.*

Howe, H. & Simkin, M. G. 2006. Factors Affecting the Ability to Detect Spreadsheet Errors. *Decision Sciences Journal of Innovative Education,* 4**,** 101-122.

Jones, D. M. 2008. *The New C Standard (Identifiers) An Economic and Cultural Commentary.*

Keller, D. 1990. A Guide to Natural Naming. *SIGPLAN notices,* 25**,** 95-102.

McKeever, R. & McDaid, K. 2010. How do Range Names Hinder Novice Debugging Performance? *EuSpRIG.* London.

McKeever, R., McDaid, K. & Bishop, B. 2009. An Exploratory Analysis of the Impact of Named Ranges on the Debugging Performance of Novice Users. *EuSpRIG.* Paris, France.

O'Beirne, P. 2005. *Spreadsheet Check and Control: 47 key practices to detect and prevent errors,* Gorey, Systems Publishing.

Panko, R. R. & Ordway, N. 2005. Sarbanes-Oxley: What About all the Spreadsheets? Controlling for Errors and Fraud in Financial Reporting. *EuSpRIG.*

Pearson, C. H. 2009. *Defined Names* [Online]. Available: http://www.cpearson.com/Excel/named.htm [Accessed November 25 2008].

Powell, S. G., Baker, K. R. & Lawson, B. 2007. *Errors in Operational Spreadsheets* [Online]. Available: http://mba.tuck.dartmouth.edu/pages/faculty/steve.powell/publications.html [Accessed].

Purser, M. & Chadwick, D. 2006. Does an awareness of differing types of spreadsheet errors aid end-users in identifying spreadsheets errors? *EuSpRIG.*

Scaffidi, C., Shaw, M. & Myers, B. 2005. *The "55M End-User Programmers" Estimate Revisited.* Pittsburgh,: Institute for Software Research Human-Computer Interaction Institute.