The Role of Visualization Tools in Spreadsheet Error Correction from a Cognitive Fit Perspective

Suparna Goswami
National University of Singapore, sups.ghanvatkar@gmail.com

Hock C. Chan
National University of Singapore, chanhc@comp.nus.edu.sg

Hee W. Kim
National University of Singapore, kimhw@comp.nus.edu.sg

Follow this and additional works at: https://aisel.aisnet.org/jais

The Role of Visualization Tools in Spreadsheet Error Correction from a Cognitive Fit Perspective*

Suparna Goswami
Department of Information Systems
National University of Singapore
suparnag@comp.nus.edu.sg

Hock Chuan Chan
Department of Information Systems
National University of Singapore
chanhc@comp.nus.edu.sg

Hee Woong Kim
Department of Information Systems
National University of Singapore
kimhw@comp.nus.edu.sg

Abstract

Errors in spreadsheets pose a serious problem for organizations and academics. This has resulted in ongoing efforts to devise measures for reducing errors or efficient ways of correcting them. Visualization tools are often advertised as means for improving spreadsheet error correction performance. This study investigates the role of visualization tools in spreadsheet error correction. For this purpose, this study proposes a framework for classifying activities associated with spreadsheet error correction. The framework is to highlight the activities that are important for correcting different types of spreadsheet errors, and to show how different visualization tools can aid error correction by effectively supporting these activities. By identifying "chaining" as one of the most important activities from the framework, this study uses cognitive fit theory to examine the effects of a visualization tool that supports chaining on spreadsheet error correction performance. Experimental methodology is used to test the outcome of cognitive fit between the error correction task and the visualization tool. The results of the experiment highlight the importance of cognitive fit between the type of task and the visualization tool for attaining better performance. This study also provides guidelines for designing and developing tools for spreadsheet error correction.

Keywords: *Spreadsheet, Visualization Tools, Activity Framework, Cognitive Fit*

The Role of Visualization Tools in Spreadsheet Error Correction from a Cognitive Fit Perspective

1. Introduction

After nearly three decades of spreadsheet use throughout organizations and various business functions, errors in spreadsheets are still common and often lead to serious consequences (Ayalew et al., 2000; Chan, 2004; Clermont et al., 2002; Panko, 2005; Panko and Halverson, 1996; 2001). A summary of reported spreadsheet audits since 1995 indicates that 94 percent of the audited spreadsheets had errors, with an average error rate of 5.2 cells per hundred cells (Panko, 2006). The critical nature of spreadsheet errors and their consequences is underlined by the fact that there are professional and academic organizations (such as the European Spreadsheet Risks Interest Groups or EuSpRIG[1]) dedicated to spreadsheet research.

The EuSpRIG website reports a number of recent occurrences of spreadsheet errors in the public domain and the consequences of such errors. For instance, in 2006, spreadsheet errors resulted in wrong tender award decisions in the Office of Government Commerce in the United Kingdom, and the Columbia Housing Authority in the United States incurred expenses amounting to more than $200,000 because of errors that overpaid landlords and allocated too much room to tenants. Spreadsheet errors have resulted in the loss of financial control, public embarrassment for organizations and individuals, legal challenges, bad press, exposure to public audit, political embarrassment, and damage to careers (EuSpRIG, 2007).

Given the severity of errors in spreadsheets and their consequences, research has been directed toward finding ways to alleviate the problem. Researchers have proposed strategies to reduce the occurrence of errors during spreadsheet development and for faster identification and correction of errors in existing spreadsheets. However, for spreadsheet users, the main method of checking and correcting spreadsheets is the same as that was practiced when spreadsheets were first introduced about three decades ago – users have to read through and analyze the textual formulas in the spreadsheet cells to identify errors. This process of error correction in spreadsheets is not only extremely tedious and difficult, but also quite error-prone (Panko, 1999; Panko and Sprague, 1998; Teo and Lee-Partridge, 1999).

Researchers have often advocated the use of visualization support or visualization tools to prevent spreadsheet errors as well as aid in their identification and correction (Chadwick et al., 2001; Chan et al., 2000; Clermont, 2003; Davis, 1996; Igarashi et al., 1998; Lentini et al., 2000; Rajalingham et al., 2000; Sajaniemi, 2000). As such, existing spreadsheet research is often based on a software engineering perspective, with the primary objective of designing and developing tools that can help prevent errors during spreadsheet development or make them easier to identify for correction later on (e.g., Burnett et al., 2002; Chadwick et al., 2001; Hendry and Green, 1993). However, most of the suggested visualization tools were developed as research prototypes and lacked empirical support and validation, and hence have not gained widespread acceptance (Chan, 2004).

In the information systems (IS) literature, spreadsheet research has primarily focused on identifying ways of improving user performance in spreadsheet error correction, and findings have suggested that domain knowledge and spreadsheet expertise can affect performance in spreadsheet error correction (Galletta et al., 1993), and that different representations of the spreadsheet can result in differing error correction performance (Galletta et al., 1996; 1997). Other research (Chan et al., 2000; Chan, 2004) has investigated the role of visualization tools in spreadsheet error correction and found that visualization tools lead to better performance for certain kinds of errors, but not necessarily for all error types. While previous IS research indicates that error correction performance in spreadsheets can depend on the presentation format, or the visualization tool used, a theoretical explanation as to why and how this happens is lacking. This often makes it difficult to interpret and extend the findings in a coherent manner for future research and practice. Consequently, spreadsheet research has received relatively sparse attention in the IS research arena, even though there has been no decline in the incidence of errors in spreadsheets (Panko, 2005).

---

[1] EuSpRIG (http://www.eusprig.org/index.htm)

Therefore, the objective of this paper is to examine the effects of visualization tools on spreadsheet error correction. A theoretically grounded knowledge of how visualization tools support spreadsheet error correction could help researchers gain a better understanding of the phenomenon and assist software developers in designing and developing more effective spreadsheet visualization tools. As a theoretical basis, this study adopts cognitive fit theory (Vessey, 1991; Vessey and Galletta, 1991) and the extended cognitive fit model (Sinha and Vessey, 1992), in particular, to explain how a visualization tool supports spreadsheet error correction by helping to achieve cognitive fit between the tool and the task. It starts with analyzing *spreadsheet error correction*[2] to understand the activities involved in the process. We propose an activity framework for classification of the spreadsheet error correction activities. The main purpose of the framework is to help identify different kinds of visualization tools that can be targeted toward each activity in the framework.

We select chaining activity — one of the core activities of the framework — for further analysis. Chaining is considered one of the most important activities involved in spreadsheet error correction (Ballinger et al., 2003; Chan et al., 2000; Igarashi et al., 1998; Shiozawa et al., 1999; Sajaniemi, 2000). Using a prototype visualization tool developed for this research and laboratory experiment as the research methodology, this study empirically investigates the effect of visualization support on the time users take to correct spreadsheet errors. This study thus aims to advance theoretical understanding of the role of visualization tools in spreadsheet error correction and to offer practical suggestions for developing and upgrading visualization tools.

The text is organized as follows: The next section constitutes the conceptual body of this paper and elaborates on the activity framework, chaining activity, how visualization tools can be used to support the chaining activity, and the effect of visually-supported chaining activity on error correction performance. Cognitive fit theory (Vessey, 1991; Vessey and Galletta, 1991) is used to develop hypotheses for the study. We then describe the research methodology, analyze the experimental results, and discuss the theoretical and practical implications of the study along with its limitations. The paper concludes with a summary of contributions.

2. Conceptual Background

2.1. A Framework for Spreadsheet Error Correction

In order to examine the effect of visualization tools on spreadsheet error correction, we first analyze the process and the activities involved in spreadsheet error correction. For this purpose, this study adopts the work of Ellis (1989), who identified six major characteristics or patterns in information seeking: *starting, chaining, browsing, differentiating, monitoring,* and *extracting*. These are the activities involved in information search and retrieval. This framework of activities has been applied extensively to explain general literature search activities for different groups of users, such as social science researchers (Ellis, 1989), engineers, and research scientists in an industrial setting (Ellis and Haugan, 1997). It has also been used to provide recommendations for the design of interactive information retrieval systems (Belkin et al., 1995), and for the experimental evaluation of such information retrieval systems (Borlund, 2000). Choo et al. (1999) showed that the activities involved in web browsing are very similar to those in information search and retrieval, and that the original activity framework can be extended to web browsing.

In spreadsheet error correction, users typically are given a spreadsheet in which they have to search for, identify, and correct errors. Searching is an important aspect of spreadsheet error correction, and there are likely to be similarities between the activities involved in spreadsheet error correction and the six major activities involved in other forms of information seeking. For instance, during a web search, a user begins by identifying websites that contain information of interest; similarly, *starting* error correction in a spreadsheet typically involves the user browsing through a large number of cells to identify those that are likely to have errors and require a closer check. In literature searches as well as web searches, after identifying the sources with information of interest, a user has to follow up on references or links to find related materials or websites. Similarly, for spreadsheet error correction,

---

[2] The term *spreadsheet error correction* is used to refer to the act of correcting spreadsheet errors. It includes understanding the structure of the spreadsheet, identifying the errors and correcting them.

after identifying cells that call for closer inspection, the user has to follow cell references in the formulas to identify related cells, i.e., perform the *chaining* activity.

Browsing in the context of spreadsheet error correction typically constitutes a general undirected scanning of cells. Based on the nature of the spreadsheet and the purpose that it serves, the user might have to *differentiate* cells based on their properties or the roles they play in formulas. For example, some cells can be input cells, some can be output cells, and some can play the role of decision variables in formulas. *Monitoring* involves keeping track of the changes made in the spreadsheet and their impact on various cells. In the context of spreadsheet correction, monitoring involves observing the effect of corrections made in the existing formulas or cell values on other spreadsheet cells. To be able to understand the structure of a spreadsheet and form logical units of groups of cells, users have to *extract* information from the spreadsheet cells. Thus, we find that for each activity from the activity framework for information seeking, there is a parallel activity in the domain of spreadsheet error correction. Hence, Ellis' (1989) activity framework can be adapted and applied to spreadsheet error correction. The first row of Table 1 shows the activities from the original framework, while the second row shows the corresponding activities associated with spreadsheet error correction.

Table 1: An Activity Framework for Spreadsheet Error Correction						
Activities →	**Starting**	**Chaining**	**Browsing**	**Differentiating**	**Monitoring**	**Extracting**
Literature search related moves (Ellis 1989)	Identifying sources of interest	Following up on references in the given material	Scanning tables of contents or headings	Assessing or restricting information according to its usefulness	Receiving regular reports from selected sources	Systematic-ally working a source to identify material of interest
Spreadsheet error correction	Identifying cells in the spread-sheet for closer check	Following cell references in formulas	General (undirected) scanning of cells in the spread-sheet	Assessing and differentiating cells based on their properties or roles	Monitoring updates during the correction process	Extracting information from cells to form bigger logical units

Note: First row is adapted from Choo et al. (1999)

This study is the first attempt to systematically categorize and delineate spreadsheet error correction activities by proposing the activity framework. The framework serves as a structure to classify spreadsheet error correction activities and identify different kinds of visualization tools that can be targeted toward each activity. To provide some initial validation of the usefulness of the framework, a few existing visualization tools are reviewed and mapped to the activities of the framework. Being able to do this indicates that prior research and commercial developments in spreadsheets already have directly or indirectly taken into account some of the activities of the framework. The visualization tools outlined here are more for indicative purposes rather than to provide an exhaustive list of all available tools.

- *Starting:* When users begin error correction on a spreadsheet, it can be difficult to decide where to start. Some visualization tools aim to help users identify interesting cells to start with. For instance, a visualization tool suggested by Chan et al. (2000) highlights cells within a selected spreadsheet area that do not follow the standard guideline of adopting a top-down and left-right structure for spreadsheet design. Another tool (Chan et al., 2000) points out cells within a selected spreadsheet area that either are not connected to any other cells or are end points in a spreadsheet model. When auditing spreadsheets, these cells are important because they are either the bottom line values in the given spreadsheet or can be error points that need to be connected further.

- *Chaining:* A spreadsheet cell can contain a formula that refers to other cells. For example, cell A1 may have a formula "=B1+C1" that refers to cells B1 and C1. The chaining activity involves identifying the precedent or dependent cells of any given spreadsheet cell. A visualization tool that

supports chaining is Microsoft Excel's Precedent and Dependent Arrow Drawing tool. The Fluid Visualization tool suggested by Igarashi et al. (1998) also supports chaining. The 'transient local view' component of this tool visualizes the dataflow structures associated with individual cells, whereas the 'static global view' visualizes the entire dataflow structure of the spreadsheet (Igarashi et al., 1998).

- *Browsing:* Users often spend a substantial amount of time simply browsing through or scanning cells without a definite aim. Tools that support this activity include Microsoft Excel's functions for adjusting display size, row width, and column height to fit some data into a screen for easier browsing.

- *Differentiating:* Typically involves the classification of cells based on certain properties. For example, a visualization tool in Chan et al. (2000) colors cells according to the input-process-output model. The CogMap tool by Hendry and Green (1993; 1994) helps in visually describing or expressing the overall structure of a spreadsheet using different tags, links, and colors. Davis (1996) describes a data dependency tool that shows cells as inputs, outputs, decision variables, parameters, or formulas.

- *Monitoring:* During the process of error correction in spreadsheets, users may want to keep track of cells that already have been tested, the changes made in spreadsheet cells, and the impact of these changes on other cells. For example, the 'What You See Is What You Test' (WYSIWYT) methodology of spreadsheet testing suggested by Burnett et al. (2002) helps users monitor the cells that they have already checked, as well as monitor the effect of corrections on other affected cells.

- *Extracting:* Some visualization tools aid in extracting information from the original spreadsheet cells. The S2 visualization tool by Sajaniemi (2000) highlights the dataflow of spreadsheets between areas containing formulas and identifies groups of cells that seem to form logical entities. The highlighted areas can extract information from the overall spreadsheet by describing the plan structure of the spreadsheet.

2.2. Chaining in Spreadsheet Error Correction

A primary difficulty with spreadsheet error correction is that of understanding the structure of the spreadsheet (i.e., the underlying relationships between various cells and how computations are done). Drawing from the proposed activity framework, *chaining* activity helps users understand spreadsheet formulas by following cell references in formulas. Accordingly, *chaining* has been considered one of the most important activities involved in comprehending a given spreadsheet model and, consequently, chaining has received a lot of research attention (Hendry and Green, 1994; Davis, 1996; Shiozawa et al., 1999; Sajaniemi, 2000; Chan et al., 2000).

The numerous cells in a typical spreadsheet model form a web with interconnections that are determined by formulas (Chan, 2004). The formulas are in text form, and are usually hidden from the view of users. Even when the formulas are displayed, it is difficult for users to trace cell connections and remember them. However, these connections are essential to understanding a spreadsheet model, as users might need to track down the meaning of cell references to understand the formulas (Hendry and Green, 1994), especially when spreadsheet errors are due to erroneous formulas. Therefore, tools are required to support *chaining*. Such tools should be able to highlight the relationships among cells that are connected through formulas.

Spatial overviews of complex data help in understanding trends, patterns, and relationships in the data (Vessey 2006). Therefore, a spatial representation of interconnections among cells in a spreadsheet can help in understanding the overall structure of the spreadsheet by emphasizing the relationships among cells. Spreadsheet formulas and computations are usually not random and tend to follow some structural rules or patterns (Tukiainen and Sajaniemi, 1996). Tools that help to create visualizations of the underlying structure of a given spreadsheet can highlight anomalies within the overall structure and help to identify potential errors. Thus, visually supported *chaining activity* can make the process of correcting spreadsheet errors simpler and more efficient, particularly when the errors are in the formulas.

The most common form of visualization support for chaining is found in Microsoft Excel, where a tool can draw an arrow from a precedent cell to its dependent cell. The dependent cell is the cell containing a formula that refers to one or more precedent cells. For instance, if cell A1 contains the formula "=B1+C1", then A1 is a dependent cell of its precedent cells B1 and C1. Intuitively, it seems that users will be faster in grasping that there is a relationship between A1 and B1 when there is an arrow going from B1 to A1, compared to when they have to read the formula in cell A1 and subsequently search for cell B1.

2.3. Role of Visualization Tools in Spreadsheet Error Correction

Spreadsheets are relatively easy to build, but their correctness is more difficult to ascertain because spreadsheet cells often contain computations and formulas that remain hidden from the user (Panko, 2006). Because the structures of computations are difficult to grasp, errors go unnoticed. Thus, mechanisms for making spreadsheets easier to comprehend and errors easier to detect are required for spreadsheet error correction (Sajaniemi, 2000). Tools can be used to draw a user's attention to relevant spreadsheet areas or cells and, therefore, help in the correction of spreadsheet errors.

By providing an overview of complex data sets, visualization generally allows for the exploitation of human spatial/visual capabilities (Tegarden,1999), and helps in identifying areas of interest, structures, patterns, trends, and relationships in data (Grinstein and Ward, 2001; Vessey 2006). Effective visualization through the use of illustration (such as images, diagrams, and graphs) can help guide a user's attention toward the key items in the presented information (Mayer, 1989). Visual search research also suggests that human attention is allocated selectively to the objects in the visual field (Lachman et al., 1979; Van der Heijden, 1992), and the ability to draw attention depends on the salience of the visual objects.

Thus, visualization tools can be used to give prominence to certain regions of the spreadsheet or to outline the overall structure of the spreadsheet. In the context of spreadsheets, visualization tools have already been developed for several purposes such as spreadsheet comprehension, error detection, debugging, and documenting, with different tools providing different amounts of user assistance (Sajaniemi, 2000).

Given the apparent difficulties in identifying and correcting spreadsheet errors, there is an urgent need to evaluate the effects of various visualization tools and to understand how they can help users in spreadsheet error correction. Such knowledge can guide future research, as well as inform spreadsheet developers and users. In this study, we use cognitive fit theory (Vessey, 1991; Vessey and Galletta, 1991) as the underlying theoretical basis to explain the role of visualization tools in spreadsheet error correction.

3. Theoretical Background and Hypotheses

3.1. Cognitive Fit Theory

Cognitive fit theory (Vessey, 1991; Vessey and Galletta, 1991) proposes that problem solving is a result of an individual's mental representation, which, in turn, depends on the problem-solving task that the user has to perform and the problem representation. Mental representation is defined as the way a problem is represented in human working memory. Problem representation refers to the way in which problem-related information is presented to the user. The problem-solving task corresponds to the actual task that the decision-maker is expected to perform. When the type of information emphasized by the problem representation matches the type of information required by the problem-solving task, there is a match between the problem representation and the problem-solving task, and the decision-maker relies on similar mental processes to act on the problem representation and to complete the task. This match results in a consistent mental representation of the problem and facilitates problem solving. However, when a mismatch occurs between the problem representation and the task, individuals have to use different cognitive processes to act on the problem representation and to solve the problem. This mismatch requires a transformation of the mental representation to derive a solution to the problem, which results in performance deterioration (Vessey, 1991).

The concept of cognitive fit in problem solving has been applied to investigate user performance in different kinds of decision-making contexts such as data presentation formats (graphs versus tables), comparison between tables and maps in managerial problem solving, different structures of maps in geographical information systems, multi-attribute data presentation, comparison between different accounting model representations, and programming tasks across different programming languages (refer to Vessey, 2006 for a detailed review on cognitive fit theory and the areas in which it has been applied).

The cognitive fit model in its basic form proposes improvements in performance due to a match between the problem-solving task and the problem representation. However, in a practical problem-solving situation, users are usually provided with tools to aid them in the problem-solving process, but the problem presentation is kept constant. Thus the cognitive fit model was extended to include the use of support tools and techniques in carrying out the problem-solving task (Sinha and Vessey, 1992). This extended cognitive fit model (see Figure 1) contends that when there is a match between the type of information required by the problem-solving task, the type of information emphasized by the problem representation, and the type of information emphasized by a given problem-solving tool, problem solving will be more effective and efficient (Sinha and Vessey, 1992). An earlier experiment investigating the performance of tools used for specifying processes in structured analysis also indicated that the usefulness of a tool must be considered in the context of the task that users are required to perform (Vessey and Weber, 1986), thus supporting the above proposition. Sinha and Vessey (1992) further found that for users performing programming tasks, irrespective of the problem representation, a match between the problem-solving tool and the task resulted in better performance.

Spreadsheet error correction is a form of problem solving, and, therefore, the notion of cognitive fit can be applied to investigate user performance in spreadsheet error correction. We are interested in understanding the role of visualization tools that support spreadsheet error correction. Thus, the extended cognitive fit model incorporating problem-solving tools is more appropriate in the context of this study.
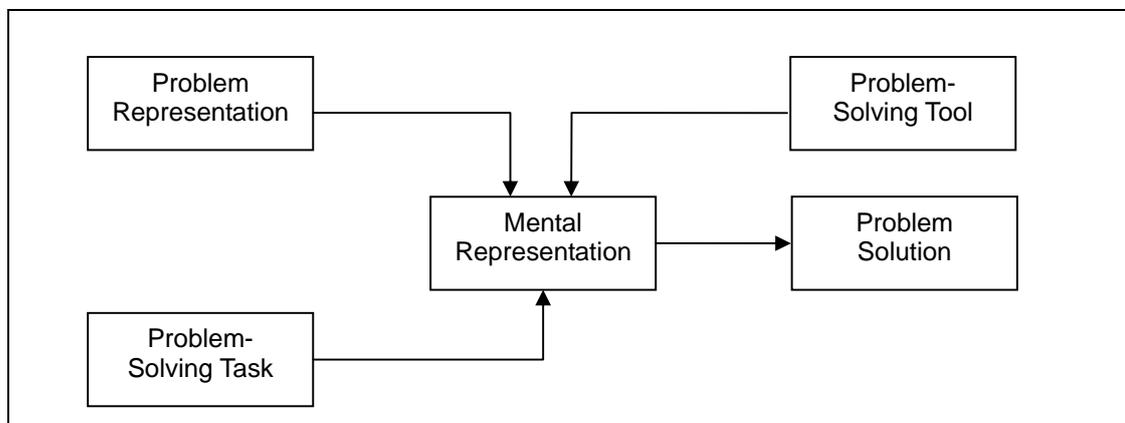

Figure 1. Extended Cognitive Fit Model (Sinha and Vessey, 1992)

While most prior empirical studies in cognitive fit theory have examined the effects of a match between task characteristics and problem representation (Dennis and Carte, 1998; Dunn and Grabski, 2001; Sinha and Vessey, 1992; Smelcer and Carmel, 1997; Umanath and Vessey, 1994; Vessey, 1994; Vessey and Galletta, 1991), this study examines the performance effects of a match between task characteristics and tool characteristics, keeping the characteristics of the problem representation as invariant. In the context of spreadsheet error correction, the problem representation takes the form of the spreadsheet models users have to work with. For the experiment in this study, we present the same spreadsheet models to all subjects. We describe the details regarding the problem-solving task and the tool in the following sections. The section ends with a discussion on the role of cognitive fit between the task and the tool.

Problem-Solving Task: Correcting Link and Non-link Errors

Subjects were required to perform the task of spreadsheet error correction. Spreadsheet errors can be of many different types, and prior research has proposed many different schemes to classify errors. It is important to distinguish between different types of errors, as they might require different types of correction methods (Teo and Tan, 1999) and consequently different tools to aid in their correction.

Spreadsheet errors have been classified as domain errors and device errors (Galletta et al., 1993); quantitative errors or qualitative errors (Teo and Tan, 1997, 1999); and, mechanical, omission, and logic errors (Panko and Halverson, 1996). However, these error classification schemes are less relevant when a user has to work on a pre-built spreadsheet model with the purpose of correcting errors, because these classification schemes are mainly meant for explaining how spreadsheet errors occur. It might not be possible to say whether an error is a mechanical error or logic error or omission error after it has already been committed. Thus, from the point of view of designing visualization tools for spreadsheet error correction, these error classification schemes are less relevant.

We use an error taxonomy proposed by Chan (2004) especially for the purpose of spreadsheet error correction. According to this taxonomy, cells with computations or formulas can contain two primary types of errors – the cell has a missing or wrong reference (*link error*), or it has a missing or wrong operator (*non-link error*). Based on this classification scheme, if the correct formula is "=C4*C5," but it is entered as "=C3*C5" instead, then "C3" is the link error. If the entry is "=C4+C5," then "+" is the non-link error. The main reason for selecting this taxonomy is that the error classification is based on formulas in the spreadsheet cells, and to some extent, understanding and correcting spreadsheet formulas depends on the chaining activity that helps establish the interconnections among cells joined through formulas. Thus, we focus on these two types of error to investigate the role of chaining and the visualization tools supporting it.

Subjects were required to perform two separate error correction tasks (this includes looking through the spreadsheet, identifying the error as well as correcting it) – a) correcting link errors and b) correcting non-link errors in the given spreadsheet models. To identify a link error, a user first has to understand the underlying structure formed by the cell connections of the spreadsheet. Thus, the user first needs to perform chaining (i.e., track down the cell references in spreadsheet formulas). Identification of link errors in the spreadsheet is typically a two-part process – understanding the relationships or associations among the cells and finding cells that have erroneous associations based on the overall structure of the spreadsheet. Because spreadsheet formulas typically tend to follow some structural rules or patterns (Tukiainen and Sajaniemi, 1996), anomalies or deviations in the overall pattern can help users to recognize and correct link errors.

On the other hand, identification of non-link errors does not require understanding the association among different cells. Rather, additional information on the purpose of the spreadsheet and how it originates might be necessary for correcting non-link errors. Chaining is thus not an essential step for the identification of non-link errors.

Problem-Solving Tool: Visualization Tool

We consider two different problem-solving tools in this study. One problem-solving tool is a standard Excel spreadsheet without a visualization tool[3]. The other problem-solving tool is a standard spreadsheet with the visualization tool. The visualization tool helps in performing the chaining activity and is similar to Microsoft Excel's arrow tool that allows users to trace chain paths. However, the visualization tool used in this study allows for a more detailed display of inter-cell relationships, as the chain paths involving all the precedents and or dependents of one particular row or column can be displayed at a click, as indicated by Figure 2, with a spreadsheet model used by Galletta et al. (1997). We provide detailed description of the tool in the Research Methodology section.

---

[3] For referencing convenience, we also refer to the visualization tool as "the tool" because the standard Excel spreadsheet is common for all subjects.

	A	B	C	D	E	F	G
1	Cash Budget		Assume:	2%	inflation/semester		
2				Fall	Spring	Summer	Overall
3	Cash, beginning			$1,000	$1,360	$673	
4	Outflows - School costs			$4,468	$4,474	$4,480	$13,422
5	Living costs			$4,172	$4,213	$4,485	$12,870
6	Inflows - Loans			$3,000	$3,000	$3,000	$9,000
7	Support from home			$6,000	$5,000	$6,000	$17,000
8	Cash, end			$1,360	$673	$700	$2,741

Figure 2. Sample Usage of the Developed Visualization Tool

The tool provides a spatial overview of the relationships among cells in the spreadsheet by using arrows to show all cells that are linked to a selected row or column of cells. Providing such a spatial overview eliminates the need to scan through individual formulas in each cell. The use of spatial representation (such as graphs and diagrams) is usually considered beneficial for the performance of tasks that involve perceptual processes, such as understanding relationships between pieces of information (Umanath and Vessey, 1994; Vessey, 2006; Xu et al., 2004).

In general, it can be assumed that computations in spreadsheets are not random, but follow certain rules of "good structure", that is, constant values and formulas are usually located in such a way that they form a harmonious structure (Sajaniemi, 2000; Tukiainen and Sajaniemi, 1996). Any formula violating this general pattern or structure can be a potential error within the spreadsheet. Therefore, the arrows used in the visualization tool also facilitate selective attention direction to a limited area in the spreadsheet i.e., localization of focus.

Using the given visualization tool localization of focus happens as follows: The user can select a particular column of the spreadsheet and use the visualization tool to view the precedent cells of the selected column. Assuming that the given spreadsheet is designed to have a good structure, it is likely that the tool will show a number of parallel arrows pointing from some other column to the selected column. If one of the cells in the selected column does not have an arrow pointing at it, or has an arrow that is not parallel to all the other arrows, then this anomaly in the overall pattern of arrows is likely to immediately grab the user's attention, resulting in localization of focus. Localization of focus is an important aspect of cognitive fit that can result in better performance, as the human attention span is a scarce commodity (Agarwal et al., 1996; Dunn and Grabski, 2001).

Cognitive Fit between Error Correction and Visualization Tools

Based on the extended cognitive fit model, a match between the characteristics of the problem-solving task and the problem-solving tool will lead to better task performance when the problem representation is invariant. Efficiency or time taken to complete the task is a commonly used and relevant measure of performance in the literature (e.g., Galletta et al., 1997; Umanath and Vessey, 1994). Accuracy is another common measure of performance for research on spreadsheet, and other areas (Chan 2004; Galletta et al. 1997; Teo and Lee-Partridge, 1999; Teo et al. 2006). We focused on time taken to correct errors as the dependent variable of interest. At the same time, we designed the experiment in such a way that there was no trade-off between accuracy and efficiency in task performance (refer to the Experiment Design section).

Another likely consequence of the match between a problem-solving tool and task is the reduction of cognitive difficulties faced by users performing the task, which results in higher levels of confidence regarding their overall task performance (Dunn and Grabski, 2001). However, past research has been unable to consistently confirm this relationship, with findings showing that lower performance was associated with higher levels of confidence among users (Dickson et al., 1977). Dunn and Grabski (2001) suggested the necessity of examining task completion time and confidence with respect to cognitive fit in a more comprehensive manner. Thus, time taken and confidence are the two dependent variables considered in this study.

In summary, this study considers two different problem-solving tasks: (1) correcting link errors and (2) correcting non-link errors. The correction of link errors depends on understanding cell connections by looking up the formula in the formula-containing cells to identify wrong or missing cell references. Thus, for correcting link errors, users need to perform two activities: understand the associations among cells, and then identify erroneous associations based on this understanding. The visualization tool can help accomplish both by presenting a spatial overview of the spreadsheet structure and by offering localization of focus. In contrast, when users are given a spreadsheet without the visualization tool, they must read and analyze the textual formulas in the spreadsheet cells to identify errors. This approach is tedious and time consuming. Therefore, the visualization tool helps users complete the task of correcting link errors faster than those who do not have the tool. Hence, it is hypothesized,

> **H1:** *Users working with the visualization tool will correct link errors faster than users without the tool.*

Cognitive fit theory suggests that a match among the problem representation, the requirements of the problem-solving task, and the problem-solving tool will alleviate the cognitive difficulties users face in performing the task. Such alleviation of cognitive difficulties can result in a higher level of confidence among users regarding their performance (Dunn and Grabski 2001). Because there is cognitive fit when users correct link errors with the visualization tool as opposed to users correcting link errors without the visualization tool, it is hypothesized,

> **H2:** *Users working with the visualization tool will have higher levels of confidence in correcting link errors than users without the tool.*

The visualization tool facilitates the perceptual processes involved in performing chaining, which is essential for the identification and correction of link errors. The identification and correction of non-link errors is more of an analytical process, since it involves understanding the purpose and content of the formulas, and then scrutinizing them for missing or wrong operators. Therefore, perceptual processes complemented by the tool are not necessary for the identification and correction of non-link errors. Identification of non-link errors may require some additional information on the purpose of the spreadsheet, such as domain knowledge (e.g., accounting knowledge if it is an accounting spreadsheet). Chaining activity is likely to be less important for non-link errors, where looking up cell references in formulas may tell little about a wrong or missing operator unless the user has additional information regarding the purpose or nature of the formulas.

Therefore, there will be a lack of cognitive fit for non-link error identification and correction with the given visualization tool. Because cognitive fit results in improved performance (Sinha and Vessey, 1992; Vessey, 1991; Vessey, 2006; Vessey and Galletta, 1991), the visualization tool would be more effective in reducing the time taken to correct link errors than in reducing the time taken to correct non-link errors. This leads to the following hypothesis:

> **H3:** *Users working with the visualization tool will correct link errors faster than non-link errors.*

When the information emphasized by the problem-solving tool (i.e., the visualization tool) matches the information emphasized by the task (i.e., correction of link errors), users are not required to perform any mental transformation of the information provided to derive a solution, and this is likely to reduce the cognitive difficulties faced in completing the task. When cognitive difficulties faced by users in performing a task decrease due to cognitive fit, users are likely to have higher levels of confidence in their task performance (Dunn and Grabski, 2001). As discussed above, while there is a fit when correcting link errors with the visualization tool, there is a lack of cognitive fit when correcting non-link errors with the visualization tool. Accordingly, it is hypothesized,

> **H4:** *Users working with the visualization tool will be more confident in correcting link errors than non-link errors.*

4. Research Methodology

4.1. Experiment Design

A laboratory experiment was performed for hypotheses testing. The problem representation was constant, but we made a distinction between the two groups of subjects by providing the visualization tool to one group only. Both groups performed the same task of identifying and correcting link errors and non-link errors. Thus, the experiment used a mixed factorial design with two factors – type of tool (spreadsheet with or without the visualization tool), and type of error correction (link error correction and non-link error correction). Availability of the visualization tool was a between-subjects factor, and type of error correction was a within-subjects factor.

The group without the visualization tool consisted of 21 subjects, and the group with the tool consisted of 18 subjects. Volunteers from first- and second-year computer science (CS) and information systems (IS) undergraduate classes having some basic working knowledge of Microsoft Excel enrolled in a major university in Singapore were the subjects for the experiment. Each subject was paid SGD$30 (about US$20) to take part in the experiment. Figure 3 shows the experiment design. Hypotheses 1 and 2 were tested by comparing the performance in cells A and B. Hypotheses 3 and 4 were tested by comparing performance in cells A and C. While the relationships between the performance values in cell D and those in the other cells are not hypothesized, cell D can be used to indicate whether the treatment is effective. For example, the difference in performance observed by comparing the values in cell A and C should not be there when comparing the corresponding values in cells B and D to indicate that the visualization tool makes a significant difference in performance.

	Link Error Correction	Non-link Error Correction
With Visualization Tool	A	C
Without Visualization Tool	B	D

Figure 3. Experiment Design

Subjects were given sufficient time to complete the given tasks. There were no monetary rewards for completing the task as fast as possible to prevent subjects from rushing through their tasks to earn more money. Instead, subjects were provided with an additional monetary incentive (another S$20) to correctly perform the task in order to ensure that subjects did not sacrifice accuracy over efficiency. Moreover, subjects could leave the experiment location as soon as they completed the experiment. Thus, the experiment was designed in such a way that subjects would be motivated to perform the task efficiently and at the same time be accurate enough to earn the additional reward. A post-hoc analysis of the number of errors corrected showed that an average of 14.56 (out of a maximum 16) errors were corrected. As the accuracy data distribution is not normal, a non-parametric test (Mann Whitney Test) across the two subject groups (i.e., with the tool and without the tool) shows a non-significant difference ($z=.958$, $p=.0338$). Therefore, the desired experimental design of having subjects focused on maximizing the accuracy of errors corrected was achieved.

4.2. Visualization Tool

A survey of the available dataflow visualization tools indicates that there are essentially two types of tools. The first type (e.g., the visualization tool in Chan (2004)) enables users to view all the links in a given worksheet simultaneously, while the second type (e.g., Microsoft Excel's trace precedent and/or dependent tools) allows users to trace the dataflow paths of one cell at a time. The first type of tool can yield a rather messy worksheet, especially for large commercial spreadsheets containing many computations, and can cause information overload for users. With the second type of tool, it may be difficult to select the first cell to trace, especially when the user is unfamiliar with the spreadsheet

being examined or when the user has to work with a large spreadsheet. Moreover, it becomes time consuming and tedious to use the tool one cell at a time.

To overcome these shortcomings, we developed a visualization tool for the purpose of this study. This tool allows the user to display the precedents and/or dependents for a whole row or column at a time. Figure 2 shows a sample screenshot of a student budget spreadsheet model (based on the model used in Galletta et al., 1997), with dataflow visualization using the visualization tool. Using the given tool, users can display all the cell linkages involving the precedents and/or dependents of a particular row or column. In Figure 2, all the precedent and dependent cells for the selected 8th row are joined by arrows.

4.3. Experimental Procedure and Task

Subjects were required to correct four spreadsheet models, each spreadsheet having two link and two non-link errors. The spreadsheet models used in the experiment are: (1) a student budget, (2) an organizational budget, (3) a project estimate, and (4) an income forecast. The models were adapted from or similar to spreadsheet models used in other empirical studies such as Panko and Halverson (1996) and Galletta et al. (1997). For illustration, the project estimate model is given in the Appendix. We developed a computer program for the purpose of randomly seeding errors into the spreadsheets, and each subject received spreadsheets with randomly incorporated errors. The main reason for using randomly generated errors was that the position of the error in the spreadsheet could affect the error correction performance. For example, it might be possible that errors that are located more toward the top-left corner of the spreadsheet are easier to detect than errors that are in the middle or bottom-right of the spreadsheet. Randomizations "assist in averaging-out the effects of extraneous factors that may be present" (Montgomery, 1997, p. 13). After identifying and correcting each error, subjects were polled on their level of confidence regarding their performance on the given task.

We controlled factors that were not part of the experimental design to provide a common ground for comparison between the two groups. For example, all subjects were tested in the same laboratory and on personal computers (PC) having the same configuration to ensure homogeneity in environmental conditions. All subjects used the same four spreadsheets. We compared subjects' prior experience in using spreadsheets across groups and found no significant difference. The only difference between the two groups was the absence or presence of the visualization tool.

We carried out the experiment over multiple sessions, and in each experimental session subjects belonged only to either the treatment group (the group working with the visualization tool) or the control group (the group working without the visualization tool). For the experiment, each subject was assigned to one PC terminal in the laboratory, and asked to read a Microsoft Excel note and a training guide. The Excel note was a one-page list of basic Excel functions (e.g., SUM and MIN) and their brief descriptions. The training guide contained a brief introduction on the different types of errors to be found in the spreadsheets, followed by a sample spreadsheet exercise. The treatment group was also provided with a small write-up on the use of the visualization tool to correct errors. A basic demonstration based on the sample spreadsheet showed all subjects how to correct errors using the user interface in the experimental software. For the treatment group, the demonstration also included a short introduction to the visualization tool. After the demonstration, students were asked to work through the sample exercise from the training guide.

After the training, all subjects proceeded through the four spreadsheet models in the same sequence. For each spreadsheet model, subjects also received a textual description of the model. The subjects were informed that there were four errors (two link errors and two non-link errors) in each spreadsheet. The experiment lasted a total of two hours and 30 minutes, giving subjects more than sufficient time to complete the tasks. In a pilot experiment, four subjects (having levels of prior experience in using Excel similar to those of the control and treatment groups) performing the same task took slightly less than two hours.

After the introductory training and demonstration, subjects had to open each spreadsheet file, correct the errors, indicate the level of confidence on their performance after each error correction, and close the file after correcting the errors. Whenever a subject opened a new spreadsheet, the program would remind him to read the corresponding description of the spreadsheet model in the training

guide booklet before proceeding. The timer started after the subject clicked on the "OK" button to confirm that he had completed reading the instructions. Immediately after the correction of each error on the spreadsheet, a pop-up window appeared in which subjects were asked to rate their confidence regarding the accuracy of their error correction on a 7-point Likert scale with (1) meaning very low confidence and (7) meaning very high confidence. For each error, the time (in minutes) was measured from clicking on "OK" until the arrival of the pop-up window just after the error had been corrected. The measured time for each error is the total time a subject took to identify and correct the error. In this study, the time reported includes the total time subjects took to perform all the sub-activities involved in spreadsheet error correction such as browsing the spreadsheet to understand its purpose, detecting errors, and then correcting the errors. Based on this conceptualization, this study examines the effect of the visualization tool on error correction performance as a whole. Although no time limits were imposed for correcting individual spreadsheets, subjects were not allowed to return to a spreadsheet once they had finished and closed it.

5. Data Analysis and Results

Table 2 reports the descriptive statistics of the experiment. The values for time taken are derived by summing the time taken to identify and correct each type of error across the four spreadsheet models for all subjects and then calculating the average value for each subject. Thus, the value 36.6 minutes in Table 2 denotes the average time taken by each subject in the treatment group (with visualization tool) to identify and correct a total of eight link errors (two errors each in four spreadsheets). For confidence, the maximum value can be 56 (7 point scale x 8 errors), and the value 52.2 represents the average confidence reported by each subject in the treatment group in identifying and correcting eight link errors.

Table 2: Descriptive Statistics		Time taken (in min.)		Confidence*	
Effect		**Link Error**	**Non-Link Error**	**Link Error**	**Non-Link Error**
With Visualization Tool	Mean	36.6	54.3	52.2	46.0
	N	18	18	18	18
	Std. Deviation	12.3	16.0	5.8	9.9
Without Visualization Tool	Mean	46.0	43.7	50.0	48.8
	N	21	21	21	21
	Std. Deviation	14.0	13.0	8.4	8.3
Total	Mean	41.7	48.6	51.0	47.5
	N	39	39	39	39
	Std. Deviation	13.9	15.3	7.3	9.1

  Note: * Confidence data is based on 7-point scale

Figures 4 and 5 are provided for easy comprehension of the experimental results. Statistical tests to decide the significance of the differences are reported below.

We performed a multivariate analysis with repeated measures using SPSS to identify the main and interaction effects of the two factors on the two dependent variables – time taken and confidence (see Table 3). The results indicate that while there is no main effect due to group (i.e., the tool), there is a main effect from error-type, and also a significant effect due to the interaction between error-type and group. When there is an interaction effect, it overrides the presence or absence of main effects.

We performed univariate tests to individually examine the effects of the two factors and their interaction on each dependent variable (see Table 4). Table 4 shows that there is a significant interaction effect of group and error type on both time and confidence.
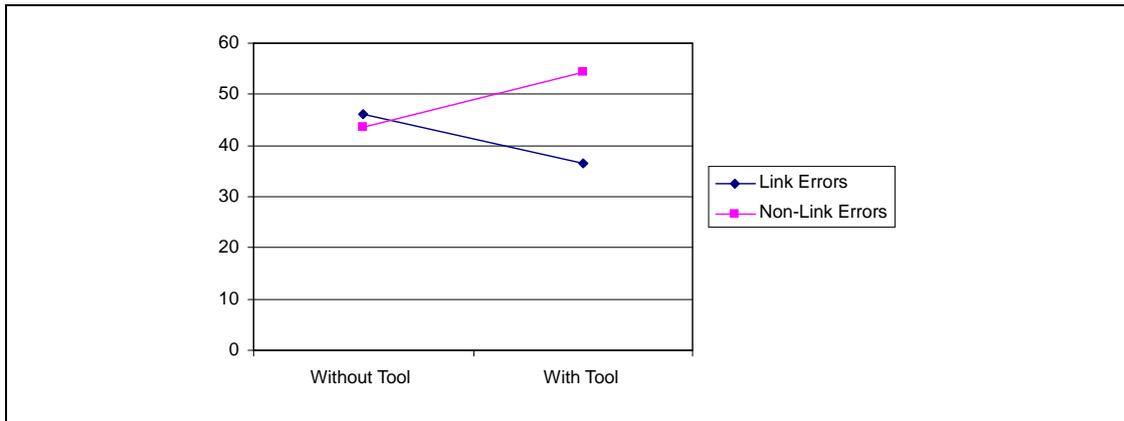
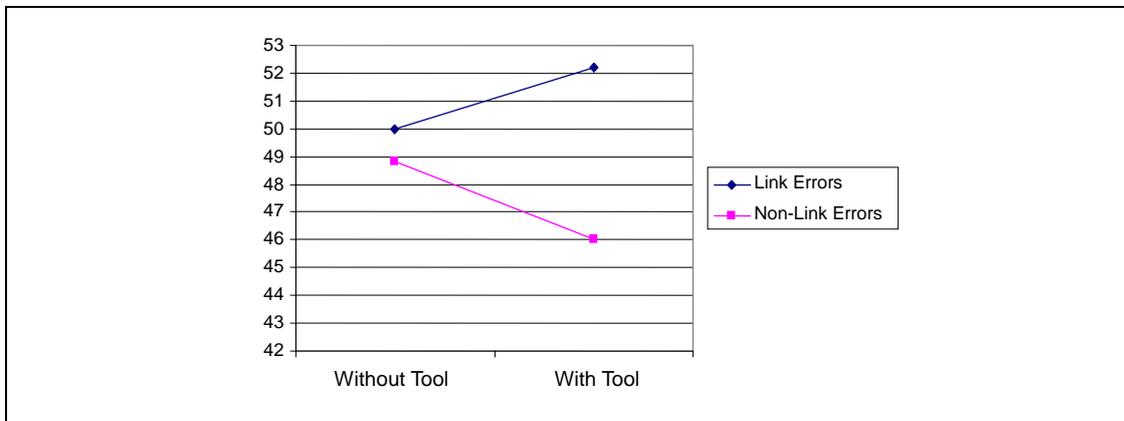**Figure 4. Effect of the Visualization Tool on Time Taken to Correct Errors**

Figure 5. Effect of the Visualization Tool on Confidence

Table 3: Multivariate Analysis Results					
Effect		**F**	**Hypothesis df**	**Error df**	**Sig.**
Between Subjects	**Group**	.012	2	36	.988
Within Subject	**Error-type**	9.51	2	36	.000
	Error-type * Group	13.46	2	36	.000

Table 4: Univariate Analysis for Time and Confidence			
Source	**Measure**	**F**	**Sig.**
Group	**Time**	0.02	.888
	Confidence	0.017	.896
Error-type	**Time**	16.0	.000
	Confidence	13.4	.001
Error-type * Group	**Time**	27.6	.000
	Confidence	5.74	.022

We carried out separate t-tests to confirm the significance of the differences among the descriptive statistics, and analyze the effects of the tool and error type on the two dependent variables. The

following descriptions focus on total effects. The time taken to correct link errors for the with visualization tool group (mean = 36.6 minutes) is significantly less that the time taken by the without visualization tool group (mean = 46.0 minutes), as indicated by the t-test (t=-2.227, p=0.032), thus supporting Hypothesis 1. For confidence in correcting link errors, a comparison of the means of the with visualization tool group (mean = 52.2) and the without visualization tool group (mean = 50.0) indicates that the presence of the visualization tool did not make any significant difference (t=.901, p=0.37). Thus, Hypothesis 2 is not supported. For the group using the visualization tool, a comparison between time taken to correct link errors (mean = 36.6 minutes) and time taken to correct non-link errors (mean = 54.3 minutes) shows that subjects took a significantly shorter time to correct link errors (t=-6.09, p=0.001), thus supporting Hypothesis 3. Similarly, for the group using the visualization tool, subjects reported significantly higher levels of confidence in correcting link errors (mean = 52.2) than in correcting non-link errors (mean = 46.0) as indicated by the t-test (t=3.48, p=0.003), thus supporting Hypothesis 4.

Other than the hypothesized relationships, we observe that for the group without the visualization tool, the difference in terms of time taken to correct link errors (mean = 46.0 minutes) and the time taken to correct non-link errors (mean = 43.7 minutes) is insignificant (t=0.95, p=0.36). Similarly, for the without visualization group, the difference in the confidence in correcting link errors (mean = 50.0) and the confidence in correcting non-link errors (mean = 48.8) is insignificant (t=1.14, p=0.27). These insignificant differences in time and confidence in correcting the two types of errors without a visualization tool further confirms that the visualization tool makes a significant difference in subjects' performance in the correction of link and non-link errors.

6. Discussion and Implications

6.1. Discussion of Findings

The results of the experiment support the basic proposition that a visualization tool designed to support chaining activity has a positive effect on the performance of tasks that are dependent on chaining, such as the correction of link errors. Results indicate that subjects using a visualization tool that facilitates chaining were significantly faster in identifying and correcting link errors, compared to subjects without the tool. Thus, the experiment provides empirical validation of the cognitive fit theory and the extended cognitive fit model (Sinha and Vessey, 1992; Vessey, 1991; Vessey and Galletta, 1991) by showing that a fit between the problem-solving task and the problem-solving tool, when the problem representation is kept invariant, results in better task performance.

However, for the second dependent variable – confidence, the experiment results show that the presence of the visualization tool did not make any significant difference for correcting link errors. Thus, cognitive fit does not necessarily result in higher levels of confidence. While Hypothesis 2 is not supported, the finding is in line with prior research, which could not confirm the relationship between cognitive fit and confidence (Dunn and Grabski, 2001). This finding could be due to the fact that unlike the measure of time as a performance indicator, confidence is a subjective measure that can be influenced by several other factors, such as an individual's personality and the complexity of the task.

Also, subjects with the tool took significantly less time (Hypothesis 3) and reported significantly higher confidence (Hypothesis 4) in correcting link errors than in correcting non-link errors. This is because, unlike link errors, non-link errors are not dependent on chaining; and the presence of the tool did not result in cognitive fit. Thus, there was a lack of fit between the visualization tool and the task of non-link error correction, while there was cognitive fit between the tool and the task of link error correction. These findings are consistent with the basic proposition of the cognitive fit theory.

In addition to the hypothesized relationships, there are some interesting observations. The finding that there was no significant difference in terms of time taken and confidence levels between link and non-link errors for the group without the visualization tool, while the difference was significant for the group with the visualization tool, further confirmed that the visualization tool had a significant influence on link error correction performance. For non-link errors, a significant deterioration of performance (both in terms of time taken and confidence reported) was observed when subjects were

provided with the visualization tool. There can be two possible explanations for this deterioration of performance.

The first explanation can be that subjects were distracted by the visualization when it was not required. This explanation is based on the notion of information overload and its effect on human beings. It has been shown that when users are faced with an increased information overload (when the information is redundant and does not convey any additional meaning with respect to the task), they will attempt to use all the available information; this results in increased cognitive effort to solve the task (Umanath and Vessey, 1994). In the current experiment, subjects who worked with the visualization tool had access to information (in the form of arrows depicting cell relationships) that was unrelated to the task of identifying non-link errors. This unrelated information could have resulted in their getting distracted by the arrows, which were not helpful in correcting non-link errors, and, therefore, they ended up spending more time in identifying non-link errors and being less confident about their performance.

The second explanation is based on the notion of "localization of focus" (Agarwal et al., 1996; Dunn and Grabski, 2001), which can be applied to understand the underlying mechanism that subjects work through on a given spreadsheet to identify and correct errors. In the absence of a visualization aid, there was nothing to guide subjects; thus their attention did not get drawn to any specific area of the spreadsheet. It was likely that they began with a random scanning of the whole spreadsheet and corrected errors as and when they were able to detect them, and they ended up spending an almost equal amount of time in correcting both types of error. However, the visualization tool showed the precedent and dependent relationships, and thus helped in highlighting areas where there were deviations from the standard patterns or relationships among the cells. Localization of focus to areas with link errors is thus likely to have happened in the presence of the visualization tool. So subjects were able to identify and correct link errors first, and then try to find the non-link errors, and therefore took more time correcting non-link errors than correcting link errors.

6.2. Limitations

The results of this study should be interpreted in the context of its limitations. The use of a controlled laboratory experiment gives rise to limitations that are inherent to this research methodology. The use of student subjects is a practical limitation; however, prior research has also argued that this does not necessarily affect the generalizability of the results (Campbell, 1986; Dipboye and Flanagan, 1979). Using CS/IS majors as subjects may raise concerns that these students are more experienced in using Excel spreadsheets compared to a general population of students. However, the error correction tasks that subjects were required to perform did not require any specialized Excel skills and so we do not view this as a matter of major concern, as all subjects had similar levels of experience with Excel.

Another possible limitation of this study arises from subjects in the treatment group being introduced to the visualization tool for the first time during the experiment. Since they had to work through four spreadsheets, each containing four errors, two of each type, it is possible that subjects got more familiar with the visualization tool over time, and this could have had an effect on their performance as they progressed from one spreadsheet to the next. Future studies can be designed so as to expose subjects to the visualization tool ahead of the experiment and allow them to become familiarized with the tool over a longer period of time before performing the actual experimental tasks.

The results of this study are specific to the visualization tool used in this study. The tool shows precedents for all cells within a single row or a single column simultaneously. Microsoft Excel's trace precedents tool can show precedents for only one cell at a time; hence it may take more time for users to click on every cell, instead of an entire row or column. At the other extreme, the visualization tool used by Chan et al. (2000) allows precedents to be displayed over the entire spreadsheet. Although this can reduce clicking time, especially in the case of large spreadsheets, there is a trade-off in terms of the clarity of the representation -- too many arrows can result in a messy spreadsheet and overload users with information. Therefore, more empirical studies may be needed to determine the optimal number of cells, rows, or columns that can be selected at a time to display the precedent

cells using the visualization tool. This will balance time spent clicking and messiness of the spreadsheet.

Finally, this study is primarily targeted toward chaining, which is believed to be a very important activity in spreadsheet correction. However, spreadsheet error correction involves other activities as well, such as starting, browsing, and differentiating. For testing other activities in the framework, hypotheses and empirical evaluations involving different kinds of support tools and different error types are needed. Future research can examine other activities in the framework and the effects of supporting tools on error correction. Further, the current tool could support other activities in addition to chaining. For instance, the visualization tool is likely to help users in deciding where to start looking for errors (i.e, starting). However, this study only considers the role of the visualization tool in supporting chaining.

6.3. Theoretical Implications

This research has several implications for theory. First, this study draws on Ellis' (1989) framework for classifying information-seeking behavior and shows that it can be used for characterizing spreadsheet error correction activities. The framework explains the different activities involved in spreadsheet error correction, and has considerable implications for future spreadsheet research. The framework allows existing visualization tools to be categorized in terms of the activities they support, thus making it easier to determine their roles in spreadsheet auditing and error correction; researchers can also devise specific ways of validating their effectiveness. In other words, the framework allows researchers to link tools to activities and investigate the corresponding implications for performance.

Visualization tools are often seen as a way to address the problems faced in detecting and correcting errors in spreadsheets; and consequently, several types of visual support tools have been proposed in the literature (Burnett et al., 2002; Chan et al., 2000; Hendry and Green, 1994; Igarashi et al., 1998; and Sajaniemi, 2000). However, until now there has been little theoretical basis to support or validate claims regarding the usefulness or effectiveness of these tools. Thus, the second important implication of this study is to provide a theoretically motivated understanding of the role of visualization tools in spreadsheet error correction.

We use cognitive fit theory (Sinha and Vessey, 1992; Vessey, 1991; Vessey and Galletta, 1991) as the theoretical basis to explain the effect of a visualization tool on spreadsheet error correction. This study is the first of its kind to apply cognitive fit theory to spreadsheet error correction, and suggests that better performance can result when there is cognitive fit between the visualization tool and the error correction task. Thus, another important implication of this study is the opening up of a new domain of application for cognitive fit theory (Vessey, 1991; Vessey and Galletta, 1991), and more specifically the extended cognitive fit model (Sinha and Vessey, 1992), which is more suitable in the context of this study.

Most previous research has hypothesized directly from visualization tool to performance, without providing any theoretical reasoning behind the relationships and often reporting insignificant effects or mixed findings (e.g., Davis, 1996; Chan, 2004). This study can help in interpreting the findings of earlier studies. Davis (1996) reported an experiment that compared the arrow tool (which gives a visual display of arrows from precedent cells to dependent cells) with the method of listing precedents and dependents textually as found in older versions of Excel (version 3.0). The experiment did not find any significant difference in error correction performance. Applying cognitive fit perspective and the activity framework, an explanation for this finding is that the errors that subjects had to correct did not emphasize the chaining activity that was supported by the provided arrow tool. Therefore, the mental processes involved in solving the experimental task in Davis (1996) did not match the mental processes that were facilitated by visualization support from the tool.

Using a similar visualization tool, Chan (2004) reported mixed results on time taken to correct different error types. These mixed findings also reflect the basic proposition of the current study -- that spreadsheet error correction is not a single activity, but can be broken down into a number of steps or separate activities. Different tools can be targeted toward supporting different activities.

Finally, in line with cognitive fit theory, the results of this study indicate that a visualization tool is effective only when there is a match between the tool and the problem-solving task at hand. Performance enhancement due to the visualization tool was restricted to link errors, which are reliant on chaining activity, as opposed to non-link errors. Thus, compared to previous studies (such as Davis, 1996; Teo and Lee-Partridge, 1999; Teo and Tan, 1997), this study contributes to a finer understanding of the effect of visualization tools on spreadsheet error correction, showing the main and interaction effects.

Future research may consider using the "thinking aloud" mechanism and "protocol analysis" to get a better understanding of the process subjects go through when correcting errors in spreadsheets, both in the presence and absence of visualization tools. This can help in further validating the usefulness of visualization tools in spreadsheet error correction.

6.4. Practical Implications

Providing a visualization tool in itself is not useful if the tool is not suitable for the error correction task at hand. In fact, there is a probability that such a tool can be counter-productive, as reflected by the deterioration of performance in correcting non-link errors in the experiment. Thus, any form of visualization support for spreadsheet error correction should match the requirements of the task to be effective. In this context, the activities involved in error correction first have to be identified, and then tools targeted toward specific activities should be designed.

Various types of errors should be analyzed to identify the activities that are crucial in correcting such errors. Identifying an activity or a sequence of activities that are critical in correcting different errors is an important step toward designing visualization tools that are better matched for different errors, and can enhance performance in spreadsheet error correction. For instance, this study shows that chaining is more important for the correction of link errors than for the correction of non-link errors. Prior research has already shown that domain expertise and spreadsheet expertise have a positive effect on spreadsheet auditing performance in terms of speed and the number of errors that are correctly identified (Galletta et al., 1993). The current research additionally suggests that knowledge on the typology of spreadsheet errors, the activities involved in error correction, and the tools that can be used for performing these activities can positively affect spreadsheet error correction. Therefore, organizations can consider providing training to spreadsheet users and auditors to improve their expertise in spreadsheet error identification and correction. Such training could provide an overview of the different types of errors that can occur in spreadsheets and the different activities that are involved in spreadsheet error correction, and help in the recognition of activities that are salient in correcting each error type and an understanding of the various error correction tools that are available and the activities that they support.

The activity framework proposed in this paper contributes toward this direction. It serves the purpose of delineating various error correction activities and matching them with appropriate visualization tools. This paper provides preliminary guidelines for establishing such a match by reviewing some of the tools already in existence for spreadsheet auditing, and by tracking the different error correction activities these tools can support. As a case in point, this study also shows that for correcting link errors, chaining activity is highly important, and therefore a visualization tool that supports chaining is effective in correcting link errors. Future research can further develop these guidelines and use them to train spreadsheet users on which tool is suitable for which error, or the order in which tools from a given tool-set should be used when correcting spreadsheets. Users can improve their performance in spreadsheet error corrections by matching the error correction activity and the tools.

The proposed framework can act as a guideline for software developers in designing and developing spreadsheet error correction tools targeting each error correction activity in the framework. Software developers can consider integrating different tools and developing a rule-base that manages the relationships among error types, activities, and tools. Such a set of integrated tools with a rule-base could facilitate spreadsheet error corrections by selecting and activating the matched tool.

7. Conclusion

Spreadsheet errors are common and often have far-reaching consequences (Panko, 2005). Research, especially in the domain of software engineering or design sciences, has been advocating the use of visualization tools as an answer to the high incidence of spreadsheet errors and issues faced when auditing spreadsheets and correcting errors in them. The use of spreadsheets and the error rates in them have seen no decline over time (Panko, 2005). However IS research (except for a few examples, such as Chan, 2004; Galletta et al., 1993; Galletta et al., 1997; Panko and Halverson, 2001) has been rather silent in addressing the problem with spreadsheet errors. This research strives to address the problem by providing theoretical and practical insights toward understanding the problem of spreadsheet error correction.

Going beyond previous research, this study provides firm theoretical development and testing based on cognitive fit theory to explain the effects of a visualization tool on spreadsheet error correction with the development of an activity framework of spreadsheet error correction. The activity framework delineates and categorizes the spreadsheet error correction activities, and cognitive fit theory is used as the theoretical basis to explain the role of a visualization tool on spreadsheet error correction. The framework contributes to a better understanding of spreadsheet error correction activities and helps in establishing a match between the activities and their supporting visualization tools. By focusing on chaining activity – which is considered one of the most important activities in spreadsheet error correction -- this study highlights the importance of cognitive fit between the type of task (which in the present case is that of link error correction and non-link error correction) and the visualization tool for attaining better performance. The findings provide guidelines for designing more effective means of visual support in spreadsheet error correction.

References

Agarwal, R., Sinha, A. P., and Tanniru, M. "Cognitive Fit in Requirement Modeling: A Study of Object and Process Technologies," *Journal of Management Information Systems* (31:2), 1996, pp. 137-162.

Ayalew, Y., Clermont, M. and Mittermeir, R. T. "Detecting Errors in Spreadsheets," in *Proceedings of EuSpRIG 2000 Symposiumon Spreadsheet Risks, Audit and Development Methods,* 2000. (http://citeseer.nj.nec.com/485147.html)

Ballinger, D., Biddle, R. and Noble, J. "Spreadsheet Visualization to improve End-user Understanding," in *Proceedings of the Australian Symposium on Information Visualization* (24), 2003, pp. 99-109.

Belkin, N. J., Cool, C., Stein, A. and Thiel, U. "Cases, Scripts and Information-seeking Strategies: On the Design of Interactive Information Retrieval Systems*," Expert Systems and Applications* (9:3), 1995, pp. 379-395.

Borlund, P. "Experimental Components for the Evaluation of Interactive Information Retrieval Systems," *Journal of Documentation* (56:1), 2000, pp. 71-90.

Burnett, M., Sheretov, A., Ren, B. and Rothermel, G. "Testing Homogeneous Spreadsheet Grids with the 'What you See Is What You Test Methodology'," *IEEE Transactions on Software Engineering* (28:6), 2002, pp. 576-594.

Campbell, J. P. "Labs, Fields, and Straw Issues," in *Generalizing from Laboratory to Field Settings*, E. A. Locke (Ed.), Lexington, MA: Lexington Books, 1986, pp. 269-279.

Chadwick D., Knight, B. and Rajalingham, K. "Quality Control in Spreadsheets: A Visual Approach using Color Codings to Reduce Errors in Formulae," *Software Quality Journal* (9:2), 2001, pp. 133-143.

Chan, H. C., Ying, C. and Peh, C. B. "Strategies and Visualization Tools for Enhancing User Auditing of Spreadsheet Models," *Information and Software Technology* (42:15), 2000, pp. 1037-1043.

Chan, H. C. " Spreadsheet Visualization Effects on Error Correction," in *Proceedings of the Third Annual Workshop on HCI Research in MIS*, 2004, pp. 100-104.

Choo, C. W., Detlor, B. and Turnbull, D. "Information Seeking on the Web: An Integrated Model of Browsing and Searching," *ASIS Annual Meeting Contributed Paper,* 1999, URL: http://choo.fis.utoronto.ca/fis/respub/asis99/

Clermont, M. "Analyzing Large Spreadsheet Programs," in *10th Working Conference on Reverse Engineering,* WCRE, 2003, pp. 306-315.

Clermont M., Hanin, C. and Mittermeir, R. "A Spreadsheet Auditing Tool Evaluated in an Industrial Context," in *Proceedings of the 3rd Annual Symposium of The EuSpRIG*, 2002 (available online at http://www.sysmod.com/eusprog.02.htm).

Davis, J. S. "Tools for Spreadsheet Auditing," *International Journal of Human-Computer Studies* (45:4), 1996, pp. 429-442.

Dennis, A. R. and Carte, T. A. "Using Geographical Information Systems for Decision Making: Extending Cognitive Fit Theory to Map-based Presentations," *Information Systems Research* (9:2), 1998, pp. 194-203.

Dickson, G. W., Senn, J. A. and Chervany, N. L. "Research in Management Information Systems: The Minnesota Experiments," *Management Science* (23:9), 1977, pp. 913-923.

Dipboye, R. L., and Flanagan, M. F. "Research Settings in Industrial and Organizational Psychology: Are Findings in the Field more Generalizable than in the Laboratory?" *American Psychologist* (34), 1979, pp. 141-150.

Dunn, C. and Grabski, S. "An Investigation of Localization as an Element of Cognitive Fit in Accounting Model Representations," *Decision Sciences* (32:1), 2001, pp. 55-94.

Ellis, D. "A Behavioural Model for Information Retrieval System Design," *Journal of Information Science* (15:4/5), 1989, pp. 237-247.

Ellis, D. and Haugan, M. "Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment," *Journal of Documentation* (53:4), 1997, pp. 384-403.

EuSpRIG – *European Spreadsheet Risk Interest Group* (http://www.eusprig.org/), 2007.

Galletta, D. F., Abraham, D., El Louadi, M., Lekse, W., Pollalis, Y., and Sampler, J. "An Empirical Study of Spreadsheet Error-Finding Performance," *Accounting, Management, and Information Technology* (3:2), 1993, pp. 79-95.

Galletta, D. F., Hartzel, K. S., Johnson, S., Joseph, J. L. and Rustagi, S. "An Experimental Study of Spreadsheet Presentation and Error Detection," in *Proceedings of the 29th Annual Hawaii International Conference on System Sciences,* 1996, pp. 336-345.

Galletta, D. F., Hartzel, K. S., Johnson, S., Joseph, J. L. and Rustagi, S. "Spreadsheet Presentation and Error Detection: An Experimental Study," *Journal of Management Information Systems* (13:3), 1997, pp. 45-63.

Graham, J. M. "Interaction Effects: Their Nature and Some Post Hoc Exploration Strategies," *Annual meeting of the Southwest Educational Research Association*, Dallas, 2000.

Grinstein, G. G. and Ward, M. O. "Introduction to Data Visualization," in *Information Visualization in Data Mining and Knowledge Discovery*, U. Fayyad, G. G. Grinstein, and A. Wierse (Eds.) Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 21-45.

Hendry, D. G. and Green, T. R. G. "CogMap: A Visual Description Language for Spreadsheets," *Journal of Visual Languages and Computing* (4:1), 1993, pp. 35-54.

Hendry, D. G. and Green, T. R. G. "Creating, Comprehending, and Explaining Spreadsheets: A Cognitive Interpretation of what Discretionary Users think of the Spreadsheet Model," *International Journal of Human-Computer Studies* (40:6), 1994, pp. 1033-1065.

Igarashi, T., Mackinlay, J. D., Chang, B.-W. and Zellweger, P. "Fluid Visualization for Spreadsheet Structures," in *Proceedings of the 14th IEEE Symposium on Visual Languages*, 1998, pp. 118-125.

Lachman, R., Lachman, J. L. and Butterfield, E. C. *Cognitive Psychology and Information Processing: An Introduction*, Lawrence Erlbaum, Hillsdale, NJ, 1979.

Lentini, M., Nardi, D. and Simonetta, A. "Self-instructive Spreadsheets: An Environment for Automatic Knowledge Acquisition and Tutor Generation," *International Journal of Human Computer Studies* (52:5), 2000, pp. 775-803.

Mayer, R. E. "Models for Understanding," *Review of Educational Research* (59:1), 1989, pp. 43-64.

Montgomery, D. C. Design and Analysis of Experiments, 4th ed., 1997, John Wiley, New York.

Panko, R. R. "Applying Code Inspection to Spreadsheet Testing," *Journal of Management Information System*s (16:2), 1999, pp. 159-176.

Panko, R. R. *Spreadsheet Research Website* (http://panko.cba.hawaii.edu/ssr/), 2006.

Panko, R. R. "What We Know about Spreadsheet Errors," 2005, available online at http://panko.cba.hawaii.edu/ssr/Mypapers/whatknow.htm

Panko, R. R. and Halverson, R. P. "Spreadsheets on Trial: A Survey of Research on Spreadsheet Risks," in *Proceedings of the 29th Hawaii International Conference on System Sciences* (HICSS), 1996, pp. 326-335.

Panko, R. R. and Halverson, R. P. "An Experiment in Collaborative Development to Reduce Spreadsheet Errors," *Journal of the Association for Information Systems* (2:4), 2001, pp. 1-29.

Panko, R. R. and Sprague, Jr., R. H. "Hitting the Wall: Errors in Developing and Code Inspecting a 'Simple' Spreadsheet Model," *Decision Support Systems* (22:4), 1998, pp. 337-353.

Rajalingham, K., Chadwick, D., Knight, B. and Edwards, D. "Quality Control in Spreadsheets: A Software Engineering-Based Approach to Spreadsheet Development," in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, HICSS, 2000.

Rosnow, R. L. and Rosenthal, R. "If You're Looking at the Cell Means, You're Not Looking at Only the Interaction (Unless All Main Effects Are Zero)," *Psychological Bulletin*, (110:3), 1991, pp. 574–576.

Sajaniemi, J. "Modelling Spreadsheet Audit: A Rigorous Approach to Automatic Visualization," *Journal of Visual Languages and Computing* (11:1), 2000, pp. 49-82.

Shiozawa, H., Okada, K. and Matsushita, Y. "3D Interactive Visualization for Inter-Cell Dependencies of Spreadsheets," in *IEEE Symposium on Information Visualization*, 1999, pp. 79-82.

Sinha, A. P. and Vessey, I. "Cognitive Fit: An Empirical Study of Recursion and Iteration," *IEEE Transactions on Software Engineering* (18:5), 1992, pp. 368-379.

Smelcer, J. B. and Carmel, E. "The Effectiveness of Different Representations for Managerial Problem Solving: Comparing Tables and Maps," *Decision Sciences* (28:2), 1997, pp. 391-420.

Tegarden, D. P. "Business Information Visualization," *Communications of the AIS* (1:1), 1999.

Teo, H. H., Chan, H. C. and K. K. Wei, "Performance Effects of Formal Modeling Language Differences: A Combined Abstraction Level and Construct Complexity Analysis," *IEEE Transactions on Professional Communication*, 49(2), 2006, pp. 160-175.

Teo, T. S. H., and Lee-Partridge J.E. "Effects of Error Factors and Prior Incremental Practice on Spreadsheet Error Detection: An Experimental Study," *Omega, International Journal of Management Science* (29:5), 1999, pp. 445-456.

Teo, T. S. H., and Tan, M. "Quantitative and Qualitative Errors in Spreadsheet Development," in *Proceedings of the 30th Hawaii International Conference on System Sciences*, HICSS, 1997, pp. 149-155.

Teo, T. S. H., and Tan, M. "Spreadsheet Development and 'What-If' analysis: Quantitative versus Qualitative Errors," *Accounting, Management and Information Technology* (9:3), 1999, pp. 141-160.

Tukiainen, M. and Sajaniemi, J. "Spreadsheet Goal and Plan Catalog: Additive and Multiplicative Computational Goals and Plans in Spreadsheet Calculation," *University of Joensuu, Department of Computer Science, Technical Report, Series A, Report-A-1996-4*, 1996.

Umanath, N. and Vessey, I. "Multi-Attribute Data Presentation and Human Judgement: A Cognitive Fit Perspective," *Decision Sciences* (25:5/6), 1994, pp. 795-824.

Van der Heijden, A. H. C. *Selective Attention in Vision*, Routledge, New York, 1992.

Vessey, I. "Cognitive Fit: A Theory-Based Analysis of the Graphs versus Tables Literature," *Decision Sciences* (22:2), 1991, pp. 219-240.

Vessey, I. "The Effect of Information Presentation on Decision making: A Cost-Benefit Analysis," *Information and Management* (27), 1994, pp. 103-119.

Vessey, I., "The Theory of Cognitive Fit: One Aspect of a General Theory of Problem Solving?" in *Human-Computer Interaction and Management Information Systems: Foundations, Advances in Management Information Systems Series,* P. Zhang and D. Galletta (eds.) M.E. Sharpe, Armonk, NY, 2006, pp. 141-183.

Vessey, I. and Galletta, D. "Cognitive Fit: An Empirical Study of Information Acquisition," *Information Systems Research* (2:1), 1991, pp. 63-84.

Vessey, I. and Weber, R. "Structured Tools and Conditional Logic: An Empirical Investigation," *Communications of the ACM* (29: 1), 1986, pp. 48-57.

Xu, S., Fang, X., Brzezinski, J. and Chan, S. "Dual-Modal Presentation of Sequential Information," in *Proceedings of the Third Annual Workshop on HCI Research in MIS*, 2004, pp. 95-99.

Appendix: Sample Spreadsheet Model used in the Experiment

	A	B	C	D
1	**Bid Wall Spreadsheet Model**			
2				
3	**Labor Hours**			
4	Days Per Person	3		
5	Hours Per Day	8		
6	Hours Per Person	24		
7	Total Hours	48		
8				
9				
10	**Labor Cost**			
11	Pay per hour	$10.00		
12	Pay	$480.00		
13	Employer's pension fund Contribution	$96.00		
14				
15	Labor Cost	$576.00		
16				
17	**Total Costs and Bid**	**Brick**	Lava	Cement
18	Labor	$576.00	$576.00	$576.00
19	Materials	$480.00	$720.00	$1,200.00
20	Transportation	$250.00	$442.00	$610.00
21	Total Costs	$1,306.00	$1,738.00	$2,386.00
22	Profit Margin	30%	30%	10%
23	Markup	$391.80	$521.40	$238.60
24	Government Grant	$100.00	$100.00	$300.00
25	Bid	$1,438.02	$1,943.46	$2,092.14
26				
27	Lowest Bid:	$1,438.02		
28				

	E	F	G	H	I
3	**Wall Volume (Cubic metres)**				
4	Height (m)		6		
5	Length (m)		20		
6	Thickness (m)		2		
7	Volume (cubic metre)		240		
10	**Materials Costs**		**Brick**	Lava	Cement
11	Cost Per Cubic Metre		$2.00	$3.00	$5.00
12	Materials Costs		$480.00	$720.00	$1,200.00
14	**Transportation Costs**		**Brick**	Lava	Cement
15	Material Density (kg/cubic metre)		1	1.8	2.5
16	Truck Load (kg)		240	432	600
17	Cost Per Unit of Load ($/Kg)		1	1	1
18	Transportation Costs		$250.00	$442.00	$610.00

About the Authors

Suparna Goswami is currently a Ph.D. candidate in the Department of Information Systems at National University of Singapore. Her research interests are in the areas of organizational adoption of information technology innovations, organizational information technology strategy and human-computer interactions and spreadsheet visualizations. Her research has been published in a number of international conferences such as ICIS, AMCIS, PACIS and AOM Meeting.

Hock Chuan Chan is an associate professor at the Department of Information Systems, National University of Singapore, Singapore. He has B.A. and M.A. degrees from the University of Cambridge and a Ph.D. degree from the University of British Columbia. His research interests include user-database interaction, spreadsheet visualization, and information systems acceptance. He has been involved in the program committees of ICIS, PACIS, and HCI/MIS workshop. He is on the editorial board of *Journal of Database Management*, *Journal of Electronic Commerce Research* and *MIS Quarterly*.

Hee-Woong Kim is assistant professor in the Department of Information Systems at the National University of Singapore. His research interests include human-computer interactions, virtual community, and IS infusion. His research work has been published in *Communications of the ACM*, *Data Base, Decision Support Systems, European Journal of Operational Research, Information and Management, International Journal of Human-Computer Studies, Journal of the American Society for Information Science, Journal of the Association for Information Systems*, *and others*.