ECIS 2011 Proceedings

European Conference on Information Systems (ECIS)

10-6-2011

BIASES IN USEFULNESS ASSESSMENT: THE REALIZED VALUE OF GENERATIVE SUPPORT SYSTEMS

Wietske van Osch

Michel Avital

Orr Mendelson

BIASES IN USEFULNESS ASSESSMENT: THE REALIZED VALUE OF GENERATIVE SUPPORT SYSTEMS

van Osch, Wietske, University of Amsterdam, 1018 WB Amsterdam, The Netherlands, w.vanOsch@uva.nl

Avital, Michel, University of Amsterdam, 1018 WB Amsterdam, The Netherlands, avital@uva.nl

Mendelson, Orr, Tel Aviv University, Tel Aviv 61390, Israel, orrmendelson@gmail.com

Te'eni, Dov, Tel Aviv University, Tel Aviv 61390, Israel, teeni@tau.ac.il

Abstract

Research on information systems (IS) adoption and acceptance has frequently relied upon self-reported measures of system usefulness. In this study, we compare self-reported with computer-monitored measures of usefulness. In a series of group experiments, participants were asked to assess the usefulness of three applications—two Generativity Support applications and one Baseline application that served as a benchmark. With no exceptions, self-reported usefulness was consistently lower than computer-monitored usefulness. Although the two Generativity Support applications provided a significant added value to enhancing group performance—as demonstrated by computer-monitored measures of usefulness—groups rated these applications as less useful than the Baseline application. We explain this paradox using the Technological Frames theory to argue that the Baseline application was rated as more useful because it fitted better with the users' existing technological frames. The Generativity Support applications, however, violated users' existing technological frames and therefore were rated as less useful, despite their positive effect on group performance. These results demonstrate how anchoring can lead to misperception of usefulness that in turn may hinder the diffusion of innovation in spite of its technological advantage. Furthermore, our findings suggest that research on IS acceptance should adopt multiple measures of usefulness simultaneously and use self-reported measures with caution, in particular when evaluating new, unfamiliar systems.

Keywords: Usefulness, Judgement, Generativity Support Systems, Interaction Analysis, Technological Frames Theory, Human-Computer Interaction (HCI)

1 Introduction

To what extent does system XYZ enhance your job performance? Participants are often asked questions like this in studies of information systems (IS) adoption and acceptance. However, we know very little as to how confident we should be about the answers given by subjects when judging the usefulness of information systems. Critical studies have long suggested that there should be a concern about the validity of retrospective self-reports by users (Collopy, 1996; Hufnagel and Conca, 1994). In this paper, we compare self-reported measures of system usefulness and computer-monitored measures of system usefulness. Usefulness is a frequently used construct in IS research and self-reported usefulness measures are often used as a surrogate for the degree of satisfaction or acceptance of a system (Lee *et al.*, 2003). Results of such studies might affect the choice of organizations in adopting a particular information system or otherwise. In addition to its implication for practice, this study offers an opportunity to examine the general validity of the frequently used self-reported measures.

The objective of this paper is to assess the validity of measuring usefulness through a comparative study of two different measures of usefulness, namely self-reported usefulness and computer-monitored usefulness. In order to compare these two measures, we conducted a series of group experiments in which participants were asked to use three applications and subsequently assess their usefulness. These applications included two Generativity Support applications and one Baseline application[1]. The computer-monitored measure of usefulness showed that the Generativity Support applications were more useful (i.e. helped participants to generate more and better ideas) in comparison to the Baseline system. Nonetheless, using the self-reported measure, participants rated the Baseline system as more useful than the two Generativity Support applications. In other words, the two measures of usefulness produced contradictory results: while the *computer-monitored* measure rated the Generativity Support applications as more useful, the *self-reported* measure rated the Baseline application as more useful.

In order to address the discrepancy between the two measures of usefulness, we augmented the experiments with ethnographic and interaction analyses of video data from the three experimental sessions. In conclusion, using Technological Frames theory (Orlikowski and Gash, 1991, 1994), we explain how people's a priori assumptions, expectations and knowledge about technologies—which are influenced by prior experience using the same or similar technologies—influence their self-reported assessments of new, unfamiliar technologies. Given that the two Generativity Support applications were incongruent with the users' technological frames, these applications were perceived as less useful despite the fact that they actually enhanced the participants' generative output in concrete terms of quantity and quality of ideas.

In the next sections, following a brief summary of some relevant literature on usefulness and the validity of measures thereof, we provide a detailed description of the data collection and analysis processes. Then, we discuss empirical evidence and examine the findings in light of the Technological Frames theory. Finally, we explore issues for further research and provide recommendations regarding the application of usefulness measures in IS research.

2 Usefulness

Usefulness is one of the key constructs in studies of IS adoption and acceptance. However, popular models of technology use—such as the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB) model—have relied heavily on self-reported measures of usefulness, popularly referred to as perceived usefulness. Perceived usefulness is defined as the degree to which a person believes that using a particular system would enhance his or her job performance (Davis, 1989); i.e. a system high in perceived usefulness is one for which a user believes it has a positive use-performance relationship.

In IS research there has been little discussion about the relationship between computer-monitored, i.e. more objective, measures of usefulness and self-reported, i.e. subjective, measures of usefulness. However, there have been some studies on the relation between computer-monitored measures and self-reported measures of other aspects of IS use. For instance, in comparing computer-monitored measures of usage with self-report data, Rice and Shook (1990) concluded that computer-monitored measures may be conceptually more valid than self-assessments. However, the authors argue that the two types of measures may be valid representations of different aspects of usage. In another study, Rice (1988) argues that while computer-monitored data are empirically more *reliable* measures of

[1] The *Baseline application* is a barebones application that was used to provide a benchmark or a reference point to measuring the effect of the generativity support provided by the Visualization and Semantics applications. The Baseline application does not offer users any generativity support in relation to their specific task-context. It provides merely an interface that allows users to record and organize their ideas, like an electronic white board, and does not thereby enhance their generative capacity.

system usage than are self-reported data, they are not necessarily more *valid*. We believe this is particularly true in the context of perceived usefulness, as it is the perception of usefulness—i.e. self-reported usefulness—that can be taken as a behavioral indicator of values and preferences (Robinson, 1988) toward the technology despite the reliability of this perception. Therefore, we argue that it is wise to use and compare multiple measures for assessing usefulness simultaneously, including both self-reported and computer-monitored measures (Rice and Shook, 1990).

Lee *et al.* (2003) in their discussion of leading researchers' perspectives on TAM research, mention that these researchers suggested the investigation of actual usage and the relationship between self-reported and actual usage. We believe that a similar investigation is valuable with respect to perceived usefulness, as potential biases in the self-reported usefulness can be the source of rejection of useful information systems. Therefore, in this paper we aim to assess the validity of measuring usefulness by comparing self-reported measures of usefulness with computer-monitored measures of usefulness.

3 Research Approach

3.1 Study Settings and Context

In this study, we adopted a multi-method approach for comparing self-reported measures of usefulness with computer-monitored measures of usefulness. We augmented quantitative data from lab-controlled group experiments with qualitative data from an ethnographic and interaction analysis of video data from the group experiments.

Three group experiments aimed to test the usefulness of two Generativity Support applications vis-à-vis a Baseline application that set the benchmark. Generativity Support applications are designed to enhance the generative capacity of individuals or groups; that is, their ability to produce something ingenious or at least new in a particular context (Avital and Te'eni, 2009). In this study, we test the usefulness of three applications as follows:

- *Visualization application*—an application that offers generativity support to users by providing **images** of objects or settings that are related to their specific task-context. These images trigger new ideas or configurations by providing users with new insightful points of view thereby potentially enhancing their generative capacity.

- *Semantics application*—an application that offers users generativity support by providing **eliciting sentences** that are based on templates of solution structures that are composed with nouns and verbs taken from the textual task. These sentences trigger new configurations or possibilities by providing users with novel and unusual combinations of words, thereby potentially enhancing their generative capacity.

- *Baseline application*—a barebones application that offers users no generativity support in relation to their specific task-context. The system provides merely an interface that allows users to record and organize their ideas, like an electronic white board, and does not thereby enhance their generative capacity. The Baseline application was used to provide a benchmark or a reference point for measuring the effect of the generativity support provided by the Visualization and Semantics applications.

Given our general research interest in enhancing the generative capacity of communities, we test these applications in the context of CMMN (pronounced 'common'), which is a community for sustainable personal mobility. CMMN aims to develop a new type of electric car and to offer a revolutionary mobility concept for the future, thereby challenging society's current mobility concepts. Hereto the community uses an online collaboration platform where the members can discuss and engage in creative, intelligent and enterprising perspectives on mobility issues. Considering that generative capacity refers both to producing something ingenious as well as to challenging the status quo and transforming social reality (Avital and Te’eni, 2009), the members of CMMN community provide a good context for testing and comparing measures of usefulness in the context of Generative Support application usage.

3.2 Data Collection

The study was conducted during a "garage meeting" that we organized at the Rotterdam School of Engineering. Garage meetings are CMMN's face-to-face meetings that aim to bring together members of the community.

All in all, a group of 15 engineers, designers, students, policy makers and other car enthusiasts were present at this meeting. We randomly divided the participants into three experimental groups of five people each. Each experimental group was assigned one of three classrooms, which were all similar in design and setup. Additionally, each group was provided with one computer, a large screen, and chairs that were organized in a circle around the computer and the screen. All tables, papers and pens were removed from the rooms in order to stimulate people to work together and use the computer for executing the experimental assignment, i.e. for generating ideas. Each group was supervised by one facilitator, who would read out the instructions, time the sessions, and manage the shifts between the three stages of the experiment (see Table 1 below). The facilitator also provided a general introduction and the system included all other instructions for the assignment. Moreover, he or she was in charge of monitoring activities with the aim of assessing both the interaction and generativity dynamics.

All groups were assigned the same challenge, which required the groups to come up with solutions for developing or optimizing an electric vehicle that allows a family to travel from the Netherlands to the South of France during their summer holidays. Moreover, they were provided with two Generativity Support applications—namely *Visualization application* and *Semantics application*—as well as a barebones application, referred to as *Baseline application*—in order to generate solutions for this challenge. However, the sequence in which these three different applications were provided differed between the groups (see Table 1 below).

	Stage 1 ⇨	**Stage 2** ⇨	**Stage 3**
Group A	Baseline application	Visualization application	Semantics application
Group B	Visualization application	Semantics application	Baseline application
Group C	Semantics application	Visualization application	Baseline application

Table 1. Overview of Experimental Groups and Stages

The different sequence of applications allowed us to (a) make comparisons *between* different groups and (b) to compare *within* each group. In other words, we could compare measures of usefulness of each Generativity Support application both between and within groups. Using each of the three applications, a group had 20 minutes to generate and document their ideas. For each idea, the group had to come up with a name and a short description (2-3 sentences) in which they described the aim and implementation of the idea. The program automatically stored all the ideas that were generated and categorized them by the application with which they were generated in order to have an overview of the number of ideas generated and their content.

For each idea, the groups were also asked to rate them for value and for ease of implementation—each on a five-point Likert scale—in order to obtain a measure of quality for each idea. These measures were stored and monitored by the computer. Finally, upon completion of the last stage of the experiment, the participants were asked to fill out an individual survey in which they provided basic demographics as well as responding to a set of questions that together composed a perceived usefulness score (adapted from Adams *et al.*, 1992) per each application. During this final stage, the facilitator verified that the participants had stopped working together and filled out the survey individually.

In addition to collecting the computer-monitored data (quantity and quality of ideas) and the survey data (demographics and perceived usefulness), we also videotaped the experimental sessions allowing for multifaceted qualitative data analysis. Video data offers several advantages, in particular in the context of analyzing human-computer interaction. First, videos are a powerful tool for capturing data about how people interact with computers, since it provides a record and a sequential stream of natural observations, including subtle elements that are difficult to capture (e.g. body language) (Mackay, 1989). Second, video data preserve the context as well as the content of the experimental sessions, allowing for a contextually rich interpretation of findings.

3.3 Data Analysis

We used two different measures of usefulness, namely self-reported usefulness and computer-monitored usefulness, as follows:

- ***Computer-monitored usefulness*** as measured by:
  - *Quantity of ideas*, i.e. actual count of ideas, as stored in the computer system. The quantity of ideas is an objective measure of usefulness.
  - *Quality of ideas*, based on value and ease of implementation and adapted from Ronen and Pass, 2007 (see Appendix 1), as rated by the group itself upon generation and stored in the computer system. The quality of ideas is a subjective measure; however, it is reported immediately and not retrospectively.
- ***Self-reported usefulness*****:** as rated by each individual group member in the post-experiment survey using the perceived usefulness score as adapted from Adams *et al.*, 1992 (see Appendix 2). Self-reported usefulness is a subjective and retrospective measure.

We believe quantity and quality of ideas adequately reflect usefulness; however, it is important to keep in mind that usefulness is a more multi-faceted concept and could potentially encompass more than quantity and quality of ideas. This further implies that computer-monitored measures of usefulness will vary according to the particular system that is tested. Hence, in the context of Generativity Support we believe quantity and quality of ideas are satisfactory proxies for understanding how useful the system was for enhancing Generativity.

Building on a multi-method approach (Campbell and Fiske, 1959), we examined the concept of *usefulness* using two methods: a self-reported measure (i.e., perceived usefulness scale), and a computer-monitored measure (i.e., a nominal count of each idea and its respective quality.) After the experiments, the participants validated the computer-monitored measures of usefulness, thereby increasing our confidence in the validity of these measures.

In order to compare effectively the different measures of usefulness, we calculated index numbers[2] for each of the measures in order to obtain one general standardized metric. These standardized scores were subsequently assessed and compared *within* and *between* groups and for all groups *together*.

Given that we wanted to make comparisons between self-reported measures of usefulness and computer-monitored measures of usefulness, we also had to analyze the more subtle aspects of the activities of the individuals and the group in the experimental sessions. Hereto, we used a combination of ethnographic and interaction analysis (Suchman and Trigg, 1991) based on multiple viewings of over six hours of video data. Ethnographic analysis involves the careful study of activities and relations between activities in a complex social setting (Myers, 1999). Interaction analysis refers to the

[2] Index numbers are always calculated with respect to a base period or base state; in our case we calculated the index scores of the *Visualization* and *Semantics applications* with respect to the *Baseline application.* Index numbers are calculated by dividing the scores of the *Visualization* and *Semantics applications* by the *Baseline application.* This provides an insightful number of the percentage of change that occurs between the usefulness of the Generativity Support applications as compared to the *Baseline application*

in-depth investigation of the interaction between people with each other and with the material environment (Suchman and Trigg, 1991) – for instance, the application. Our focus was on the group interactions and the use of the applications in order to corroborate and expand the results from the experiments regarding the validity of measures of usefulness. The videos were viewed and re-viewed, transcribed and noted independently by the two researchers, generating activity and interaction logs, in order to allow for shared editing control (Mackay, 1989). Subsequently notes and transcripts were discussed and integrated by the two researchers in order to (1) identify important activities, relations and interactions in the groups, (2) gather both usual and unusual instances, and (3) juxtapose multiple analytic perspectives on the same instances. In order to maintain the richness of the video data, we added meaningful snapshots to illustrate identified activities, relations and interactions.

4 Results

In this section, we first discuss the comparison of self-reported and computer-monitored measures of usefulness based on the results of the three group experiments. Afterward, we use the findings from the video analysis to contextualize and expand the experimental results.

4.1 Experimental Results

Table 2 below shows that there are evident biases in self-reported measures of usefulness—as rated by the groups subjectively and retrospectively, when compared to computer-monitored measures of usefulness—as represented by quantity and quality of ideas. The last column in this table represents the bias between the two measurement methods, which is the ratio of self-reported usefulness over computer-monitored usefulness. This column shows that for the *Visualization* and *Semantics applications*, the self-reported usefulness—as measured through perceived usefulness—is considerably lower than the computer-monitored usefulness in all three groups, as demonstrated by all index numbers being lower than 1. In general, both the *Visualization* and the *Semantics applications* are most useful according to computer-monitored measures of usefulness, yet these two Generativity Support applications were rated as low and moderately useful by the three experimental groups. On the other hand, the *Baseline application* was rated as highly useful, despite the fact that overall it generated the fewest ideas and, as judged by the participants, the ideas were of the lowest quality in terms of value and ease of implementation.

Group	**(Stage) Application**	**COMPUTER-MONITORED USEFULNESS**			**SELF-REPORTED USEFULNESS**	**BIAS RATIO***
		Quantity stand. (raw)	**Quality stand. (raw)**	**Average stand.**	**Perceived Usefulness stand. (raw)**	
Group A	*(1) Baseline app*	*1.00* (10)	*1.00* (5.80)	***1.00***	***1.00*** (4.0)	**1.00**
	(2) Visualization app	*0.60* (6)	*0.75* (4.33)	***0.67***	***0.55*** (2.2)	**0.82**
	(3) Semantics app	*0.40* (4)	*0.98* (4.25)	***0.69***	***0.30*** (1.2)	**0.43**
Group B	*(1) Baseline app*	*1.00* (6)	*1.00* (4.67)	***1.00***	***1.00*** (3.4)	**1.00**
	(2) Visualization app	*2.17* (13)	*1.53* (7.15)	***1.84***	***0.82*** (2.8)	**0.45**
	(3) Semantics app	*2.33* (14)	*1.16* (5.43)	***1.74***	***1.00*** (3.4)	**0.57**
Group C	*(1) Baseline app*	*1.00* (9)	*1.00* (4.22)	***1.00***	***1.00*** (4.0)	**1.00**
	(2) Visualization app	*0.89* (8)	*1.13* (4.75)	***1.01***	***0.65*** (2.6)	**0.65**
	(3) Semantics app	*1.22* (11)	*1.85* (7.82)	***1.53***	***0.85*** (3.4)	**0.55**
Average	*(1) Baseline app*	*1.00* (8.33)	*1.00* (4.90)	***1.00***	***1.00*** (3.80)	**1.00**
	(2) Visualization app	*1.22* (9.00)	*1.14* (5.41)	***1.17***	***0.67*** (2.53)	**0.64**
	(3) Semantics app	*1.32* (9.67)	*1.33* (5.83)	***1.32***	***0.72*** (2.67)	**0.52**

**Bias Ratio= self-reported usefulness/ average computer-monitored usefulness*

Table 2. A comparison of self-reported and computer-monitored measures of usefulness

In the remainder of this section, we discuss the index numbers of the *Visualization* and the *Semantics applications* compared to the score of the *Baseline application.* These numbers are compared within groups, between groups, and for all groups together as illustrated in Figure 1 and 2 below.

With respect to the **within group** results, we see that for all three groups, the computer-monitored usefulness is always higher than the self-reported usefulness for both the *Visualization application* and the *Semantics application.* Despite the fact that Group A performed less well with the two Generativity Support systems than the other two groups and that their performance using the *Visualization application* and the *Semantics application* was lower than using the *Baseline application,* the index numbers show that irrespectively their self-reported usefulness was lower than the computer-monitored usefulness.

When we look at the BIAS RATIO index numbers, the last column in Table 2, we see that in Group A, the discrepancy between self-reported and computer-monitored usefulness is 18% for the *Visualization application* and 57% for the *Semantics application.* In other words, the computer-monitored usefulness is 18% and 57% higher than the self-reported usefulness for the *Visualization application* and the *Semantics application* respectively. In Group B, this discrepancy is 55% for the *Visualization application* and 43% for the *Semantics application.* In Group C, the discrepancy between self-reported and computer-monitored usefulness is 35% for the *Visualization application* and 45% for the *Semantics application.*

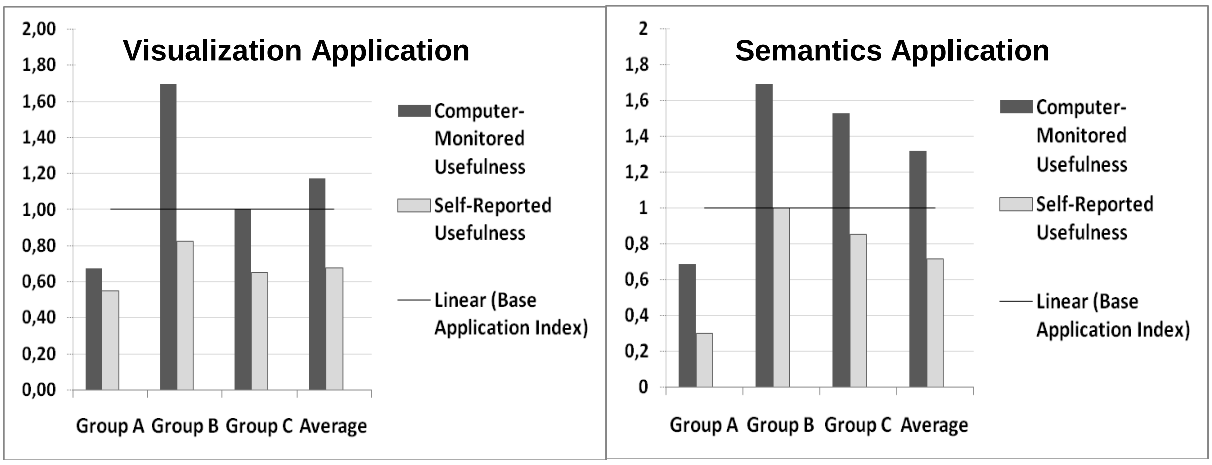

Figure 1 and Figure 2. Discrepancy between self-reported and computer-monitored usefulness

Analyzing the results **between** the three **groups**, we can see that despite individual differences in actual index numbers, the same pattern holds across the three groups. The same discrepancy between self-reported usefulness and computer-monitored usefulness therefore holds in different populations, despite Group A being an outlier—i.e. being the only group where the *Baseline application* performed better than the two Generativity Support applications.

Overall, that is, for **all groups** together the self-reported usefulness is around 36% lower than the computer-monitored usefulness for the *Visualization application* when compared to the *Baseline application*. For the *Semantics application* this discrepancy is even larger, namely 48%. This implies that on average the groups undervalued the usefulness of the *Visualization application* and *Semantics application* with 36% and 48% respectively when compared to the computer-monitored usefulness in light of the *Baseline application.*

Next, we offer in-depth insights into the interaction and generativity dynamics of the three groups based on an ethnographic and interaction analysis of the video data from the experiment. Then, in the subsequent section, we move to a theoretical discussion about why these biases in self-reported usefulness exist.

4.2 Video Analysis Results

Based on the analysis of the video data, the following will include a set of usual and unusual incidents (Mackay, 1989) from the three experimental sessions, in order to set the stage for an explanation of the biases in self-reported usefulness when compared to computer-monitored usefulness of the Generativity Support applications. For each of the three different applications—*Baseline application*, *Visualization application* and *Semantics application*—we will give a description of interactions and activities during the sessions. However, before we go into details on the dynamics of interaction and generativity, we will first give a brief description of the general ambience during the entire experiment.

Figure 3. Positive energy during experiment

In general, the mood in all groups was very positive and very comfortable (see Figure 3). The entire group process was both stimulating and pleasant at all times and did not involve pressure of any sort. Despite the fact that people had never met before, the groups started discussing and formulating ideas and solutions to the challenge—i.e. being generative—immediately. From the start groups appeared very comfortable with both the challenge and the different applications they worked with. We think this can largely be attributed to their shared culture as part of the CMMN community as well as to the structured nature of the experiment. Not only was the challenge very clearly defined, but also the process—as facilitated by the system—and the time schedule for the sessions were clearly defined. In what follows, we will see how this positive energy was evident throughout the experiment–considerably lower, however, in the *Baseline* session than in the *Visualization* and *Semantics* sessions.

The experimental session using the *Baseline application* was rather tedious in all three groups. The lower level of positive energy in the groups during this session was evident from less joking and laughing between the members of the group. Moreover, in all three groups, it seemed time went slower during this session. Whereas during the other two sessions, all three groups ran out of time, groups stopped before the 20 minutes of the *Baseline application* session was over. The interesting finding is that this also turned out to be the least generative session, in terms of quantity and quality of ideas, as shown in Figures 1 and 2 above. This might show support for the idea that humor, positive energies and vibes, and an overall positive flow are crucial for generativity (Csikszentmihalyi M., 1990; Van Osch and Avital, 2010). However, on average Group 1 performed better using this application than the other two groups, which can most likely be attributed to the sequence in which applications were used, namely that this group started with the *Baseline application* for solving the challenge (see Table 1 presented earlier in this paper).

The sessions using the *Visualization application* were characterized by a lot of excitement, given that different images gave very diverse inputs for solving the challenge and spurred a lot of jokes. Even though the images were clearly used in generating ideas, the videos reveal that groups would primarily view and discuss pictures in the first half of the session and afterward end up discussing, more than actually viewing, pictures in the course of being generative. Despite the high level of positive energy during the *Visualization* session, all three groups criticized the system as being more of a distraction than a source of inspiration. We believe that this largely influenced the groups' perception of whether the system helped in generating more and better ideas and therefore affected the moderate usefulness rating of this application.

Similar to the *Visualization* sessions, the sessions using the *Semantics application* were characterized by a lot of excitement. Despite the fact that the groups actively used the different sentences and word combinations to generate new ideas successfully, the video reveals that, primarily in Group 1, the functioning and operating of the application created some confusion. This most likely affected the low to moderate self-reported usefulness rating groups gave to the *Semantics application* based on their perception that the system did not adequately support them in generating more and better ideas. As such, a more user-friendly application could have potentially generated even better results in terms of quantity and quality of ideas as well as in terms of self-reported usefulness.

In short, it appears that the *Visualization* and *Semantics applications* were under more scrutiny than the Baseline application. Despite the positive effect of these two applications on the energy levels and generativity of the groups, the groups were not aware of this positive effect, largely due to their own perception of the negative influence—"distraction"—of the applications on the generative process, as well as some confusion over the *Semantics application*. Consequently, groups adopted a critical and antagonistic stance toward evaluating the usefulness of the two Generativity applications for generating more and better ideas.

5 Discussion

In what follows, we will explain the biases in self-reported usefulness when compared to computer-monitored usefulness through a theoretical interpretation of the findings from both the experiments and the video analysis. Subsequently, we will summarize the contributions of this study and discuss implications for future research and practice.

In order to explain the biases in self-reported usefulness, we draw upon technological frames theory. In essence, technological frames theory focuses on technology-oriented mental models—*technological frames*—which comprise the assumptions, expectation, and knowledge that people use to understand technologies (Orlikowski and Gash, 1991, 1994). These frames are powerful in that assumptions and expectations about technologies influence the choices people make regarding the subsequent use of those technologies. Although these mental models are not entirely fixed, i.e. they do evolve over time, frames are typically self-reinforcing, even to the point of rejecting knowledge or facts that do not fit existing frames of meaning.

Assumptions, expectations and knowledge about technologies are influenced by a person's prior experience using that specific technology or a related tool. Hence, when this person or a group of people is faced with an unfamiliar technology, individuals impose their frames of the familiar technology, and the technology that is most congruent with an individual's or group's frame will be perceived more positively than tools that are less congruent. Based on our findings, we suggest that the *Baseline application*—which was evaluated as most useful—resembles the tool that the group actually works with as a community. Therefore, despite the fact that this application did not provide concrete support in the generative task, congruency with participants' existing technological frames led to a higher rating on the self-reported usefulness measure. On the other hand, the *Visualization* and *Semantics applications*, which were rated as moderately useful—despite their usefulness in generating many and high-quality ideas—were the most unfamiliar tools, hence most incompatible with the participants' existing technology frames. Because these applications violated the groups' existing technological frames, the "fact" that these applications actually supported the group in conducting their tasks and solving the challenge was rejected. Consequently, these two applications were perceived as less useful, as was also clear in the videos from the slight disparagement of these applications throughout use.

In short, it appears that users' evaluations of the usefulness of applications are largely influenced by technological frames which scrutinize new applications in light of familiar applications. Our results suggest that users are likely to undervalue the usefulness of new and unfamiliar applications with about 35-50% in comparison to a familiar application. Therefore, it seems that during the initial introduction period, the more unfamiliar an application is, the lower the self-reported score of usefulness is likely to be. Consequently, whether an unfamiliar application is actually useful—in terms of its impact on group performance or other desired outcomes—seems to have little effect on users' judgement about its usefulness. These findings may also explain in part the familiar concept of resistance to change that is often encountered in the context of the implementation of new information systems.

5.1 Contributions and Future Research

The above discussion of our findings points to several important contributions. First, by revealing a discrepancy between self-reported usefulness and computer-monitored measures of usefulness, this study sheds a new light on popular concepts in IS research in regard to system adoption. Studies on IS adoption should not only look at self-reported usefulness, but also use computer-monitored measures of usefulness in order to assess whether biases exist in the former. By using a mixed method approach, this study was able to triangulate results and thereby increase the reliability of our theoretical explanation of the biases in self-reported usefulness. Therefore, we advise future research on usefulness of IS to adopt multiple measures of usefulness simultaneously, including both self-reported and computer-monitored measures.

Second, in particular within the context of testing new, unfamiliar IS, researchers as well as designers need to be more sensitive to potential biases in self-reported usefulness, given the incompatibility of these systems with users' existing technological frames. Third, if indeed biases in self-reported usefulness exist, this points to the need to develop participatory methods for communicating these biases to users and convincing them of the usefulness of new applications or systems as demonstrated by computer-monitored measures of usefulness. As emphasized by Venkatash (2003) training may be an effective driver of acceptance for users that may be less inclined to use new applications.

Fourth, this study provided preliminary support for two principles underlying the design of Generativity Support applications and thereby showed that *Visualization* and *Semantics* can help groups to solve complex challenges by triggering more ideas and ideas of higher quality through providing new insightful points of view as well as novel and unusual combinations. Therefore, our findings show that these applications are able to spur new configurations and possibilities, and hence enhance generative capacity. Future research should attempt to validate these results and test additional system features of Generativity Support applications and therewith provide additional and more comprehensive support for the value of these applications in enhancing generative capacity.

5.2 Implications

Practically, it seems that people's perception—i.e. self-reports—of the usefulness should be taken with a grain of salt by organizations or communities when considering adoption of novel technology. Self-reports of usefulness by inexperienced testers should be evaluated accordingly, or measured only after a training and adjustment period that allows them sufficient time to get familiar and comfortable with the technology under consideration.

Furthermore, these biases in self-reported usefulness present a challenge for designers to find ways to design novel applications so as to fit people's existing technological frames, making it more likely that people positively rate and subsequently adopt the particular application or system. As the video results showed, despite the usefulness of the *Visualization* and *Semantics* applications for enhancing group generativity, the applications could have been designed in a more user friendly way, thereby stimulating even more extensive engagement with these applications, further enhancing users' generative capacity, and potentially leading to more positive evaluations.

6 Conclusion

This study found large differences when self-reported measures of usefulness were compared with computer-monitored measures of usefulness in three group experiments, in which participants were asked to use three applications and subsequently assess their usefulness. These three applications included two Generativity Support applications aimed at enhancing group generativity and one Baseline application. The computer-monitored measures of usefulness showed that the two Generativity Support applications, when compared to the Baseline application, did indeed enhance users' generative capacity—as demonstrated by computer-monitored measures of quantity and quality of ideas generated. However, when the groups were asked to rate the usefulness of these three applications, the Baseline application was perceived to be more useful than the two Generativity Support applications.

Using Technological Frames theory, we have explained how people's existing assumptions, expectations and knowledge about technologies—which are influenced by prior experience using the same or similar technologies—affect their perception and subsequent evaluation of new, unfamiliar technologies. Given that the *Visualization* and *Semantics* applications were unfamiliar and hence incongruent with users' existing technological frames, these applications were perceived as less useful despite the fact that they enhanced group performance in terms of quantity and quality of ideas. The *Baseline application,* which in general performed worse in terms of computer-monitored measures of usefulness, was rated the most useful due to it being largely consistent with users' existing expectations and assumptions. These results suggest that care should be exercised in using self-reported measures of system usefulness without analyzing and comparing these with computer-monitored usefulness, and that studies on IS adoption and acceptance should employ multiple measures of usefulness simultaneously, including self-reported and computer-monitored measures. Therefore, our results provide useful insights both to those who wish to theoretically understand the relation between self-reported usefulness and computer-monitored measures of the usefulness of systems as well as for those who wish to design novel and useful information systems with a high probability of being valued and accepted by users.

References

Adams, D. A., Nelson, R.R. and Todd, P.A. (1992). Perceived Usefulness, Ease of Use and Usage of Information Technology: A Replication. MIS Quarterly, 16 (2), 227-247.

Avital, M. and Te'eni, D. (2009). From generative fit to generative capacity: exploring an emerging dimension of information systems design and task performance. Information Systems Journal, 19, 345-367.

Campbell, D.T., and FiskeD.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Collopy, F. (1996). Biases in retrospective self-reports of time use: an empirical study of computer users. Management Science, 42 (5), 758-767.

Csikszentmihalyi M. (1990). Flow: The Psychology of Optimal Experience. New York: Harper & Row

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13, 319-339.

Hufnagel, E. M. and Conca, C. (1994). User Response Data: The Potential for Errors and Biases. Information Systems Research, 5, 48-73.

Lee, Y., Kozar, K.A. and Larsen, K.R.T. (2003). The Technology Acceptance Model: Past, Present, and the Future. Communications of the AIS, 12, 752-780.

Mackay, W.E. (1989). EVA: an experimental video annotator for symbolic analysis of video data. ACM SIGHCI Bulletin, 21(2), 68-71.

Myers, M.D. (1999). Investigating information systems with ethnographic research. Communications of the AIS, 2 (23), 1-20.

Orlikowski, W. J., and Gash, D. C. (1991). Changing frames: Towards an understanding of information technology and organizational change. In Best Paper Proceedings of the 51st Annual Academy of Management Meeting, Miami Beach, FL, August.

Orlikowski, W. J., and Gash, D. C. (1994). Technological frames: Making sense of information technology in organizations. ACM Transactions on Information Systems, 12 (2), 174-207.

Orlikowski, W.J. (2007). Sociomaterial practices: Exploring technology at work. Organization Studies, 28 (9), 1433-1448.

Rice, R. E. (1988). Access to, Usage of, and Outcomes from an Electronic Messaging System. ACM Transactions on Office Information Systems, 6, 255-276.

Rice, R. E., and Shook, D.E. (1990). Relationships of Job Categories and Organizational Levels to Use of Communication Channels, Including Electronic Mail: A Meta-Analysis and Extension. Journal of Management Studies, 27, 195-229.

Robinson, J. P. (1988). Time-Diary Evidence About the Social Psychology of Everyday Life. In McGrath, J. E. (Ed.). “The Social Psychology of Time”, Newbury Park, CA, Sage Publications, 134-148.

Ronen, B., Pass, S. (2007). Focused Operations Management, Wiley.

Suchman, L., and Trigg, R. (1991). Understanding practice: Video as a medium for reflection and design. In Greenbaum, J., and Kyng, M., (Eds). “Design at Work: cooperative Design of Computer Systems”, Hilllsdael, NJ. Lawrence Erlbaum, 65-90.

Van Osch, W. and Avital, M. (2010). Generative Collectives. Proceedings of the International Conference on Information Systems (ICIS), Saint Louis, Missouri.

Appendixes

Appendix 1 - Solution's Quality Estimation Measure

For every team-generated solution to the challenge, please rate the following on a 5-point scale:
- Please estimate to what degree the suggested solution can solve the problem.
- Please estimate the ease of implementation of the suggested solution.
- Please estimate the economic value of the suggested solution.

Appendix 2 - Perceived Usefulness Measure

For each application, please rate the following on a 5-point scale:
-Do you feel the … application helped you to solve the challenge faster?
-Do you feel the … application is an effective way to deal with the challenge?
-Do you feel the ... application helped you to come up with more ideas?
-Do you feel the … application helped you to come up with better ideas?
-Do you feel the … application helped you to solve the challenge more easily?
-Do you feel the … application was useful in solving the challenge?