User evaluations of IS as surrogates for objective performance

Dale L. Goodhue, Barbara D. Klein, Salvatore T. March

Management Department, Terry College of Business, Brooks Hall, University of Georgia, Athens, GA 30602, USA
University of Michigan-Dearborn, 4901 Evergreen Road–1070 SOM, Dearborn, MI 48128-1491, USA
Information and Decision Sciences Department, 395 Hubert Humphrey Center, Carlson School of Management, University of Minnesota, 271 19th Avenue S., Minneapolis, MN 55455, USA

Received 8 September 1999; accepted 26 December 1999

Abstract

User evaluations of information systems are frequently used as measures of MIS success, since it is extremely difficult to get objective measures of system performance. However, user evaluations have been appropriately criticized as lacking a clearly articulated theoretical basis for linking them to systems effectiveness, and almost no research has been found that explicitly tests the link between user evaluations of systems and objectively measured performance. In this paper, we focus on user evaluations of task-technology fit for mandatory use systems and develop theoretical arguments for the link to individual performance. This is then empirically tested in a controlled experiment with objective performance measures and carefully validated user evaluations. Statistically significant support for the link is found for one measure of performance but not for a second. These findings are consistent with others which found that users are not necessarily accurate reporters of key constructs related to use of IS, specifically that self reporting is a poor measure of actual utilization. The possibility that user evaluations have a stronger link to performance when users receive feedback on their performance is proposed. Implications are discussed. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Management information systems; Decision support systems; Performance evaluation; Experimental design; Linear regression; Logistic regression

1. Introduction

Because objective measures of MIS success have not been developed, much research attention has been devoted to surrogate measures of success. Many IS researchers have argued that user evaluations (UE) of systems are an appropriate surrogate for MIS success, especially when system use is mandatory [10–13,15,22,23,25]. However, UE measures have been strongly criticized as lacking strong theoretical underpinnings [16,33]. In particular, Melone [26] has suggested that there is no clearly articulated theory relating UE with systems effectiveness. Furthermore, a scan of five key MIS journals from 1988 to 1996 produced almost no studies that measured both UE and objective performance. Three studies did assess both constructs and provide some indirect indication whether they might be considered surrogates. Szajna [32] found that unrealistic user expectations about a system affected its UE but did not affect objectively measured decision performance. Hunton [19] found
that the mode of participation in systems development (from mute to voice to choice) affected UE and objective performance in the same direction. Neither of these studies made any attempt to look directly at the relationship between UE and objective performance. Hunton and Price [20] measured both user evaluations and objective performance, but considered them as distinct constructs, with higher UE hypothesized to lead to higher commitment to task, and thus, to higher performance. They found a weak direct link between user evaluations and objective performance (a path coefficient of 0.29, significant at 0.05). It is accurate to say that even though much research relies upon an assumption that UE is an appropriate surrogate for objective performance, the question of the appropriateness of that assumption has been neglected.

If differences in user evaluations of systems can be shown to be reasonable surrogates for differences in performance, there are a number of important managerial implications. User evaluations could be very useful in measuring the impact of policy or system changes, implementing total quality programs that depend upon measures of results, or diagnosing problems and guiding remedial actions. From a research point of view, user evaluation of systems could provide a convenient dependent measure for many types of studies. However, if we cannot prove a link between user evaluations and objective performance, we should discourage their use for prediction purposes. Our study makes a first direct attempt to address this issue for a mandatory use situation.

2. The link between user evaluations of TTF and performance

2.1. Previous measures

As Melone pointed out, for most existing UE of IS instruments, the conceptual link between user evaluations and performance has not been clearly articulated. User evaluations have been developed based on at least two theoretical underpinnings: job satisfaction [1] and attitudes and behavior research [6] and each of these is weak in this regard.

For UE instruments based on job satisfaction, the implicit conceptual assertion has been that a happy worker will be a more productive worker. By implication, a worker happy with his information systems will also be more productive. But job satisfaction has been convincingly shown to have only a weak relationship to performance [2,21,39]. Therefore, it is doubtful that measuring whether IS makes employees happy will indicate that they are productive.

A second theoretical underpinning for UE instruments is the considerable work performed on the relationship between attitudes and behaviors. User evaluations based on attitudes/behavior theories (for example, the Technology Acceptance Model [5]) have made important contributions, but focus on predicting use rather than individual performance. Theories predicting use are obviously not applicable when use is mandatory. Further, to believe that UE instruments based on these theories will predict performance, we must make an heroic assumption that greater use implies better performance. That is clearly not always the case [30]. Goodhue and Thompson [18] argue that greater use leads to better performance only when there is high task-technology fit (TTF). Thus, it is hard to defend a strong link from UE to performance from either of these conceptual bases.

2.2. UE of task-technology fit

Goodhue [17] and Goodhue and Thompson [18] have argued for a third theoretical underpinning for UE instruments: TTF. The TTF perspective is applicable for both mandatory and voluntary use situations. It views technology as a means for a goal directed individual to carry out a task. TTF presumes that the performance impacts are dependent upon the fit between three constructs: technology characteristics, task requirements, and individual abilities. Thus, it is not the technology in isolation that affects performance.

Therefore, we should expect that any given characteristic of a technology will have different impacts on performance, depending upon the task requirements or type of user. For example, consider a database. One technology characteristic is the extent to which the database is ‘integrated’ (i.e. with consistently defined data across the whole organization). Integrated data may be advantageous if there is a task requirement for comparing information across divisions, but perhaps not if the need is for single division
information. Similarly, integrated data may be advantageous only to individuals who have been trained to use it, since its larger scope brings with it an increase in complexity.

The link between TTF and performance draws on two theoretical perspectives:

1. Organizational structural contingency theories [36] and the organizational information processing research [4,14,27,35] typically propose that better performance comes when the design characteristics of the organization ‘fit’ the requirements of the tasks and the capabilities of the individuals.

2. While organizational contingency theories typically focus on the organization or the subunit level, cognitive cost/benefit research applies similar reasoning at the individual level. It asserts that individuals consider the costs (mental effort of information acquisition and computation) and the benefits (impact on accuracy, speed, justifiability, etc.) before choosing an information acquisition strategy, and that these costs and benefits vary depending upon the task [3,29].

Implicit in this cognitive cost/benefit perspective is the assumption that individuals are aware of the costs and benefits of different options and make their choices on that basis. It could be argued this is a very complex task that may tax human abilities. For example, Connolly and Thorn [3] found that human decision makers do not respond optimally to costs and benefits when choosing how much information to access. On the other hand, we argue that their study demonstrates that humans do respond (though imperfectly) to cost and benefit differences and that they learn and improve their choices over time.

2.3. UE of TTF and performance

The argument for the link between UE of TTF and performance has three parts.

2.3.1. TTF affects performance

When a technology has exactly the functionality needed to complete the required actions for the task, better performance should result. Similarly, when an individual has appropriate knowledge and experience required to use the technology, better performance should result. The clearest demonstration of the link between TTF and performance at the individual level comes from MIS research on graphs versus tables. This research found little support for its original contention that graphical display of information by itself should lead to positive performance impacts. However, both Jarvenpaa [24] and Vessey [37] and Vessey and Galletta [38] made sense of the previously inconclusive findings by arguing that ‘incongruence’ or a lack-of-fit between task demands and technology characteristics should slow decision processes or lead to greater error or both. In particular, Vessey argued that some tasks require recognizing spatial relationships between variables and are best supported by graphs, while other tasks require exact values and are best supported by tables. Mismatches between data representations and tasks make it necessary for the user to do additional mental translations leading to greater cognitive effort, more elapsed time, and greater possibility of error. For highest performance, the information representation must ‘fit’ the demands of task. Both Jarvenpaa and Vessey found that TTF impacted performance in a mandatory use situation.

Managerial query tasks differ (among other things) on whether they require information from several sources or only one. Database technology implementations differ (among other things) on whether they are integrated or not. Modifying slightly the cognitive fit terminology of Vessey, a managerial query requires that subjects develop a mental representation of the task, including how to translate from the data representation implicit in the database’s definitions and structure to the data representation required by the problem solving task. If we assume a task whose information requirements span two sources, a non-integrated data environment needs a translation scheme and the mental representation is quite complex: it involves translating from two different data representations (data definitions and structure). We would expect lower TTF to result in longer times to complete a given task, and more errors as more complex mental processing is required.

Similarly, specific training that better prepares an individual for exactly the technology to be used should affect performance. Integrated and non-integrated data require different skills of the user. Neither is beyond the capability of most users, but familiarity with the technology should certainly improve speed. Training could also have an impact on accuracy.
H1: Higher task-technology fit results in better performance  
H1a: In a multi-division query situation, integrated data results in faster completion.  
H1b: In a multi-division query situation, integrated data results in greater accuracy.  
H1c: In a multi-division query situation, more appropriate training results in faster completion.  
H1d: In a multi-division query situation, more appropriate training results in greater accuracy.  
These assertions, at least at the macro level, border on common sense. However, they are included as specific hypotheses since they are part of the logic for the link between UE of TTF and performance.

2.3.2. Users can evaluate TTF  
Following the cognitive cost benefit research framework, we assume that users are aware of the costs and benefits of technology even when that use is mandatory, i.e. they are sensitive to relevant differences in IS characteristics and their implications on task processes. These include difficulty of operating the hardware, gaining access to data, formulating queries, etc. Users are also sensitive to the different outcome implications (differences in accuracy of solutions, completeness, time required, etc.) of different technologies. Since they are sensitive to differences in fit, we assume they can evaluate fit.  

H2: Actual differences in task-technology fit predict corresponding dimensions of user evaluations of task-technology fit, when subjects have used that technology for that task  
As part of developing an instrument to measure UE of TTF, Goodhue developed a model of the task of using organizational data to answer managerial questions. Based on this task model, he identified a number of dimensions along which technology would have to meet task demands. Using data from a survey of 259 users in nine companies, he showed that users can reliably evaluate TTF, and that characteristics of the technology, the task, and their interactions do significantly predict the relevant dimensions of user evaluations.

Suppose users are engaged in a task that requires consolidating information from two different data sources, and some are provided with integrated data while others have non-integrated data. H2 implies that users who are asked to evaluate whether the information they have used is compatible across the task domain will say that integrated data is better than non-integrated data. Similarly, users given a more appropriate degree of training should feel that there is a higher evaluation of the fit. Two of the components of TTF found by Goodhue were ‘sufficient data consistency for the task at hand’ and ‘adequate training for the task at hand’. Therefore, we restate H2 as

- H2a: With a query task requiring data from multiple sources, a treatment of ‘integrated data’ leads to higher user evaluations for ‘sufficient data consistency’ than a treatment of ‘non-integrated data’.  
- H2b: With a query task requiring data from multiple sources, a treatment ‘training on the appropriate database system’ leads to higher user evaluations for ‘adequate training for the task at hand’ than a treatment of ‘training on a different database system’.

2.3.3. UE of TTF should predict performance  
Finally, if TTF affects performance, and users will reflect TTF in their evaluations, then those evaluations should predict performance. Goodhue and Thompson have shown that user evaluations of TTF are statistically linked to perceived performance. But the link between user evaluations and objectively measured performance has not been tested.

H3: To the extent that differences in TTF affect objectively measured performance, user evaluations of those particular dimensions of TTF predict objectively measured performance  
We summarize all the above assertions in Fig. 1.

![Fig. 1. Model of task-technology fit, user evaluations, and performance](image-url)
3. Testing the links

3.1. Experimental design

To test the link between user evaluations of IS and objective performance, we designed a laboratory situation where we could manipulate TTF enough to cause changes in objective performance, and simultaneously capture measures of UE of TTF. As part of a required MIS course, 155 pairs of undergraduate business students participated in four 1-hr lab sessions learning and using a query language and a database to answer managerial questions about two different divisions of a fictitious company. The first three sessions were viewed as training and calibration sessions. The first two sessions acquainted subjects with the query language using questions pertaining to only a single division. The third session introduced questions that spanned the two divisions.

3.2. The task

The task faced by all subjects in the fourth session was to answer a series of managerial questions spanning the two divisions. For example, one of the managerial questions (shown in Appendix A) required subjects to determine, by division, the year-to-date shipping expenses for common carrier and company provided shipping. The top portion of Table 1 shows one version of the relevant account codes for this question. Completing this task requires translating divisional account codes into the categories needed to answer the question, accessing the appropriate data on the databases, and aggregating values from appropriate accounts to give answers for each of the divisions. This task was high in interdependence (a key task characteristic increasing uncertainty, according to Tushman and Nadler), since it involved comparing values across two different divisions. Thus, a technology that was capable of easily combining information from two different divisions should give a better TTF.

3.3. Independent variables

There were two experimental treatments: technology fit with task and individual fit with technology. Each had two levels. In addition, there were a number of control variables.

3.3.1. Technology fit with task

Technological fit with task was manipulated by providing some subjects with an integrated data environment across the two divisions, and others with separate non-integrated data environments for the two divisions. In the integrated data environment, all data for both divisions were maintained in a single set of three tables, and all account codes were consistently defined.

Table 1
Example of integrated vs. non-integrated account codes

<table>
<thead>
<tr>
<th>Lubbock division</th>
<th>Minneapolis division</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset of account codes for the integrated environment</td>
<td></td>
</tr>
<tr>
<td>611 Outside trucking (freight)</td>
<td>611 Outside trucking (freight)</td>
</tr>
<tr>
<td>612 Outside rail freight</td>
<td>612 Outside rail freight</td>
</tr>
<tr>
<td>613 Outside air freight</td>
<td>613 Outside air freight</td>
</tr>
<tr>
<td>621 Truck maintenance expense</td>
<td>621 Truck maintenance expense</td>
</tr>
<tr>
<td>622 Truck depreciation expense</td>
<td>622 Truck depreciation expense</td>
</tr>
<tr>
<td>623 Truck operating expense</td>
<td>623 Truck operating expense</td>
</tr>
<tr>
<td>Equivalent subset of account codes for the non-integrated environment</td>
<td></td>
</tr>
<tr>
<td>611 Freight-in common carrier</td>
<td>611 Outside trucking (freight)</td>
</tr>
<tr>
<td>612 Freight-in our trucks</td>
<td>612 Outside rail freight</td>
</tr>
<tr>
<td>620 Freight-out common carrier</td>
<td>613 Outside air freight</td>
</tr>
<tr>
<td>621 Freight-out our trucks</td>
<td>621 Truck maintenance expense</td>
</tr>
<tr>
<td>630 Freight-between, common carrier</td>
<td>622 Truck depreciation expense</td>
</tr>
<tr>
<td>631 Freight-between our trucks</td>
<td>623 Truck operating expense</td>
</tr>
</tbody>
</table>

a This is a subset of the full list of 67 account codes which subjects received, grouped by asset, liability, capital, expense, and revenue.
across the two divisions. In the non-integrated environment, each division had its own set of three tables, and students had to shift from one division’s environment to another. In addition, though the schema of the two databases was the same, account codes were defined differently and reflected different sets of managerial concerns, as shown in the lower portion of Table 1. Individual account codes had both different numbers and different meanings, so that it was not possible to map from account codes in one division to account codes in the other on a one-to-one basis. Thus, the integrated environment technology was specifically designed to need a reduced cognitive effort.

3.3.2. Individual fit with technology

Individual fit with technology was manipulated by the training sequence. All students received identical training during sessions 1 and 2. Subjects were given a standard tutorial document and lab instructors (there were five different instructors for the different sections) made some minimal introductory remarks and, especially in sessions 1 and 2, gave some limited individual assistance.

In sessions 3 and 4, the focus of the task was expanded to include two divisions. Training was manipulated by giving some subjects the same data environment (either integrated or non-integrated) in sessions 3 and 4, while giving other subjects one environment in session 3 and the other environment in session 4. Thus, in session 4 some subjects had experience on the data environment they were using, presumably had developed some familiarity and proficiency on that technology, and had better individual fit with technology. Other subjects in session 4 were encountering that data environment for the first time, and had lower individual fit with technology.

3.3.3. Control variables

There are a number of other factors besides our treatments of type of database and training history that could influence the outcomes of the experiment. To the extent possible, we controlled for these. First, there are some factors, such as programming knowledge or experience with other database systems that could have a substantial impact on user evaluations and performance. We asked each individual two questions to elicit a crude measure of familiarity with the technology.

Also, it seemed reasonable to control for the effect of the lab instructor on performance. Each of them taught at least two sections: at least one integrated section and one non-integrated section. Since there were five instructors, each presumably preparing and assisting students in a slightly different fashion, we included dummy variables for any possible effect.

Finally, factors such as innate abilities for this kind of task, motivation to work rapidly and/or accurately, pair dynamics, etc. might explain the differences between the pairs of subjects. To control for this we looked at the performance in session three (one of the training sessions). Specifically, we grouped all subjects into treatment cohorts (i.e. those that had exactly the same database, training history, and lab instructor in sessions 3 and 4), and normalized their session 3 time-to-complete and accuracy, based on their group means and standard deviations. This gave us two new control variables: the relative tendency of the pair to work fast, and the relative tendency of the pair to work accurately. We believe that these capture much of the difference in individual ability and motivation during session 4.

3.4. Pairs versus individuals

Though it might have been better to test our assertions using individuals rather than pairs, the realities of the educational setting made this impractical. The labs provided one terminal for every two students, and the philosophy of the lab was to encourage team work. The question then becomes: are the theoretical contentions equally applicable to this operationalization of the concepts? There are three general theoretical contentions: that TTF affects performance; that individuals will be able to evaluate TTF; and that UE of TTF will predict performance.

The rationale for the importance of fit between technology functionality and task needs is equally strong at the group level, as suggested by DeSanctis and Poole [9] and Zigurs and Buckland [40], and as noted in the organizational contingency literature. Given the context of the experiment, the couple may have slightly more complex goals, may be able to capitalize on ability synergy, or may be hampered by communication problems or the need for one to educate the other, but none of these seems to introduce
any perverse bias to the contention that better TTF will result in better performance. In fact, much work in organizational settings is done by groups of individuals.

We also note that the user evaluations in this experiment were done at the individual level, not the pair level. This too seems typical of user evaluations in organizations — work done in pairs and in groups would be the basis of individual level user evaluations. Finally, if TTF affects the performance of pairs, and individual users can successfully evaluate TTF, then individual UE should predict pair performance.

3.5. Measures and measurement validity

Performance was measured in two different ways: time-to-complete and accuracy. In each session the system automatically recorded the time of each student pair’s first and last query, allowing an easy and accurate measure of elapsed time for the exercise. Student pairs also turned in their completed reports containing the answers they had extracted from the data bases. Each answer was graded as either accurate or not, and students were given a percentage score across all answers. Means and standard deviations for all continuous variables are shown in Table 2.

User evaluations of TTF were measured by having each individual complete a questionnaire containing 18 evaluations about nine different aspects of the task-system fit for the system he or she had just used. These are a subset of dimensions and questions used in user evaluations of TTF by Goodhue and Goodhue and Thompson. These were modified slightly to apply more directly to the experimental domain. They are shown in Appendix B.

Averaging the two individual responses for the pair is justified as long as there is not too much disagreement between the individuals. Following an approach suggested by Tushman [34] we identified all pairs for which the variance between the two responses for any dimension was greater than the pooled variance for that dimension, with a statistical significance of 0.05. These pairs were given a missing value for that dimension, effectively removing them from consideration. These within-pair disagreements resulted in one ‘missing value’ for ‘consistency’, two for ‘accessibility’, three for ‘help’, four for ‘training’, and one for ‘system reliability’ out of 155.

Though most of these dimensions had been shown to be distinct across a sample of 24 groups in 10 companies by Goodhue, we used a more conservative approach to their measurement validity. First the scores for each student pair were averaged on a question by question basis. Then the 18 resultant question scores were factor analyzed across the full sample. Using a minimum eigenvalue of 1.0, the analysis produced four factors. However, this solution showed two of the nine targeted dimensions had measurement problems, and one of the factors combined questions for consistency across data sources with questions about technical problems with the system. Since the eigenvalue cut-off is somewhat arbitrary, we specified five factors (giving a minimum eigenvalue of 0.94). This second analysis was essentially the same as the first, except that questions for consistency and technical problems now loaded on separate factors. Two problematic dimensions remained: the two questions for ease-of-use and the two questions for understandability were split across two factors. These two-dimensions and their four questions were dropped and the factor analysis repeated, resulting in the same set of five consistent factors, interpreted as accessibility, consistency, training, help, and system reliability, as shown in Table 3. In particular, for the two-dimensions of particular interest (UE of consistency and UE of training), each question loaded on its factor with a coefficient of at least 0.64, and no question loaded on any other factor with higher than a 0.24 coefficient. Cronbach’s alpha for the five-dimensions are also shown in Table 3, ranging from 0.65 to 0.85. Nunnally [28] has suggested that initial construct development might utilize measures with reliabilities as low as 0.50, though later
versions of his work raised the desired figure to 0.70. For frequently used research constructs, 0.80 is recommended. The results of the factor analysis and the Cronbach’s alphas suggest that these user evaluation measures of TTF have good discriminant validity and at least marginally acceptable reliability, though the reliability for the two key dimensions of UE of TTF—consistency and training—is lower than we would like.

4. Discussion

4.1. Conclusions about the efficacy of user evaluations with mandatory use

Appendix C contains the analysis and results. The research results are summarized in Table 4. The empirical evidence supports some but not all of our hypotheses. There is good support for the contention

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Research results</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTF affects performance</td>
<td></td>
</tr>
<tr>
<td>H1a: integrated data results in faster completion</td>
<td>Supported</td>
</tr>
<tr>
<td>H1b: integrated data results in greater accuracy</td>
<td>Supported</td>
</tr>
<tr>
<td>H1c: more appropriate training results in faster completion</td>
<td>Supported</td>
</tr>
<tr>
<td>H1d: more appropriate training results in greater accuracy</td>
<td>Not supported</td>
</tr>
<tr>
<td>Users can evaluate TTF</td>
<td></td>
</tr>
<tr>
<td>H2a: UE of data consistency are influenced by data integration</td>
<td>Supported</td>
</tr>
<tr>
<td>H2b: UE of adequacy of training are influenced by appropriateness of training</td>
<td>Supported</td>
</tr>
<tr>
<td>UE of TTF predict objective performance</td>
<td></td>
</tr>
<tr>
<td>H3a: UE of data consistency predict time-to-complete</td>
<td>Supported</td>
</tr>
<tr>
<td>H3b: UE of data consistency predict accuracy</td>
<td>Not supported</td>
</tr>
<tr>
<td>H3c: UE of adequacy of training predict time-to-complete</td>
<td>Not supported, but relationship is in the proper direction and approaches statistical significance</td>
</tr>
</tbody>
</table>
that TTF affects performance (H1), and that users can evaluate the underlying TTF successfully (H2). But there is only very mixed support for the central assertion that user evaluations of TTF will predict performance (H3).

For H1, the treatment of integrated data versus non-integrated data had a definite effect on time-to-complete and accuracy. Subjects with the integrated data on average finished the task 5 min faster and were twice as likely to have 100% accuracy. Apparently, the extra cognitive effort to perform additional translations between the database and the problem data representations took more time and created more opportunity for errors. For the second treatment, subjects with training on the same database completed the task 7 min faster on average, but were not any more accurate. Apparently, the difference in familiarity with the technology could be overcome with a bit more time, but did not introduce more opportunities for errors. Possibly the different effect on accuracy from the two treatments may be due to different types of errors, with one type likely to result in explicit error messages and the other not. Errors of the type caused by unfamiliarity with the technology may have resulted in explicit error messages, while errors of the type caused by incorrect translations from one data representation to another would not cause an explicit error message. Thus, errors caused by lack of training caused delays, but were eventually corrected.

For H2, there were no surprises. User evaluations of consistency of the data sources and of training adequacy clearly reflected the manipulation in TTF that we designed into the experiment.

For H3, the evidence of user evaluations predicting performance differences is more mixed. There is moderate support for the contention that user evaluations predicted the time-to-complete measure of performance. User evaluations of data consistency were clearly significant, and evaluations of training adequacy were in the proper direction and approached significance (at 0.12).

But when considering accuracy we had not expected that people’s feelings about training would predict accuracy, since differences in training did not seem to affect accuracy. But differences in the integration of the data did clearly affect accuracy, and user evaluations of the consistency of the data clearly reflected the different treatments of integration versus non-integration. Maybe an important difference between the two types of performance in this domain was the amount of feedback received by the subjects. Time required for the task was quite obvious to the subjects, including the time required for each aspect of the task. If the lack of consistent data definitions and structure added additional time-consuming steps to the task, subjects were probably well aware of it. On the other hand, in this task domain there was no feedback on accuracy. Subjects had no way of knowing whether their answers were correct or not, until after the entire experiment and questionnaire were completed. Thus, subjects would not have had the opportunity to learn of accuracy problems, or to ascribe causes to those problems.

This suggests something a little different: our results suggest that the links are not quite strong enough, especially given the relatively large amount of variance introduced by any user evaluation. A strong link between UE and performance may require, in addition, an awareness on the part of the user of the impact of different conditions of TTF on performance. This awareness could sharpen the user’s assessment of TTF, and lower the error variance in the UE of TTF.

4.2. Limitations of this study

There are a number of limitations that should be recognized. First of all, there are two factors that should make us cautious. One is the fact that we have a very crude measure of accuracy (specifically whether or not the pair was 100% accurate) and a high (60%) proportion of the pairs who were completely accurate. It might well be that a task that resulted in more variance for the accuracy measure would surface a stronger, more significant link. The other factor is that the reliability of the two measures of UE (consistency and training) is marginal. Low reliability weakens the observable correlation between two constructs, suggesting that a higher reliability measure (a few more questions on each construct) might well result in a tighter, statistically significant link. This is particularly true for the link between UE of training and time-to-complete, which approaches significance. Though it is a bigger stretch to expect that the link from UE of consistency to accuracy
would become significant (the significance in our data is about 0.30), it is not inconceivable.

A further limitation is that the subjects were undergraduate students rather than practising managers. However, in this research we are less concerned with the specific impacts of integration and training than we are with the more general question of the impact of TTF on human performance, and the ability of humans to evaluate that TTF in a way that predicts performance.

4.3. Implications for researchers and managers

This study has investigated the question of whether user evaluations of TTF can serve as surrogates for performance in managerial decision making task domains. The empirical results fall considerably short of an unqualified endorsement. Here, it appears that the link between user evaluations and performance is strong only when there is feedback to the user on his or her performance.

These findings are in fact quite consistent with other recent findings that users are not necessarily accurate reporters for key constructs related to their use of IS. Specifically, users’ self reports of utilization may be a poor measure of actual utilization [31], and users’ assessments of the performance impacts of a system they have used may conflict with reality, even when they are given feedback [7].

It is too early to conclude that user evaluations are not a good surrogate for performance. However, our results, combined with recent indications of the limitations of user assessments of systems constructs in general, should sound a note of caution to managers looking for a measure of the impact of systems or policy changes. Even so, it is important to recognize that user evaluations may be used for political reasons even if they are not a good surrogate for performance. Clearly there needs to be more attention to understanding whether and when the link between UE and performance is strong. Though user evaluations of systems may reflect many important aspects of ‘success’, such as job satisfaction, user acceptance, etc., they should be used very cautiously as surrogates for performance, especially when feedback on performance is not available to users.

Appendix A. Sample task: analysis of freight expenses

This task requires account codes shown in Table 1. The accuracy of subjects on this task was based on the number of correct values placed in the table below for the four values of common carrier expenses and our truck expenses, for both Minneapolis and Lubock. Below is the problem as given to subjects.

Management has become concerned that its freight expenses are excessive. Freight expenses include all expenses related to the transport of goods. There are three activities involving the transport of goods: shipments from vendors to Sample Company’s warehouses, shipments to customers, and intra-company shipments between warehouses. These shipments may be done by common carrier (rail, air or truck), or they may be done using Sample Company’s own trucks.

In particular, Management wants to know what percentage of each division’s shipping is going by common carrier. For each division you must determine which accounts relate to the freight expense (including both common carrier and division trucks) by examining the account descriptions in the Chart of Accounts for that division.

Then write queries to access the data you need, switching back and forth between divisions as needed. Fill in the following report form:

---

Appendix A. Sample task: analysis of freight expenses

This task requires account codes shown in Table 1. The accuracy of subjects on this task was based on the number of correct values placed in the table below for the four values of common carrier expenses and our truck expenses, for both Minneapolis and Lubock. Below is the problem as given to subjects.

Management has become concerned that its freight expenses are excessive. Freight expenses include all expenses related to the transport of goods. There are three activities involving the transport of goods: shipments from vendors to Sample Company’s warehouses, shipments to customers, and intra-company shipments between warehouses. These shipments may be done by common carrier (rail, air or truck), or they may be done using Sample Company’s own trucks.

In particular, Management wants to know what percentage of each division’s shipping is going by common carrier. For each division you must determine which accounts relate to the freight expense (including both common carrier and division trucks) by examining the account descriptions in the Chart of Accounts for that division.

Then write queries to access the data you need, switching back and forth between divisions as needed. Fill in the following report form:

---

Appendix A. Sample task: analysis of freight expenses

This task requires account codes shown in Table 1. The accuracy of subjects on this task was based on the number of correct values placed in the table below for the four values of common carrier expenses and our truck expenses, for both Minneapolis and Lubock. Below is the problem as given to subjects.

Management has become concerned that its freight expenses are excessive. Freight expenses include all expenses related to the transport of goods. There are three activities involving the transport of goods: shipments from vendors to Sample Company’s warehouses, shipments to customers, and intra-company shipments between warehouses. These shipments may be done by common carrier (rail, air or truck), or they may be done using Sample Company’s own trucks.

In particular, Management wants to know what percentage of each division’s shipping is going by common carrier. For each division you must determine which accounts relate to the freight expense (including both common carrier and division trucks) by examining the account descriptions in the Chart of Accounts for that division.

Then write queries to access the data you need, switching back and forth between divisions as needed. Fill in the following report form:

---

Appendix A. Sample task: analysis of freight expenses

This task requires account codes shown in Table 1. The accuracy of subjects on this task was based on the number of correct values placed in the table below for the four values of common carrier expenses and our truck expenses, for both Minneapolis and Lubock. Below is the problem as given to subjects.

Management has become concerned that its freight expenses are excessive. Freight expenses include all expenses related to the transport of goods. There are three activities involving the transport of goods: shipments from vendors to Sample Company’s warehouses, shipments to customers, and intra-company shipments between warehouses. These shipments may be done by common carrier (rail, air or truck), or they may be done using Sample Company’s own trucks.

In particular, Management wants to know what percentage of each division’s shipping is going by common carrier. For each division you must determine which accounts relate to the freight expense (including both common carrier and division trucks) by examining the account descriptions in the Chart of Accounts for that division.

Then write queries to access the data you need, switching back and forth between divisions as needed. Fill in the following report form:

---

Appendix A. Sample task: analysis of freight expenses

This task requires account codes shown in Table 1. The accuracy of subjects on this task was based on the number of correct values placed in the table below for the four values of common carrier expenses and our truck expenses, for both Minneapolis and Lubock. Below is the problem as given to subjects.

Management has become concerned that its freight expenses are excessive. Freight expenses include all expenses related to the transport of goods. There are three activities involving the transport of goods: shipments from vendors to Sample Company’s warehouses, shipments to customers, and intra-company shipments between warehouses. These shipments may be done by common carrier (rail, air or truck), or they may be done using Sample Company’s own trucks.

In particular, Management wants to know what percentage of each division’s shipping is going by common carrier. For each division you must determine which accounts relate to the freight expense (including both common carrier and division trucks) by examining the account descriptions in the Chart of Accounts for that division.

Then write queries to access the data you need, switching back and forth between divisions as needed. Fill in the following report form:
Which division has a bigger portion of its shipments in common carriers? Is there much difference between the divisions?

Appendix B. User evaluation questions and dimensions — full set

Consistency: When it was necessary to compare or consolidate data from the two divisions, there were inconsistencies that made the task more difficult than it should have been.

There were times when I found that inconsistencies between the data for the two divisions made completing the task more difficult.

Training: I had sufficient training for this type of task prior to this lab session.

More training on this type of system would have been helpful before this task.

Help: The lab assistants gave me the help I needed to be able to do these tasks.

I was able to get the help I needed during the course of this lab session.

System reliability: There were unplanned technical problems that made this task more difficult than it should have been.

The system was subject to unexpected technical problems or crashes.

Accessibility: It was easy to get access to the data that I needed.

I could get data I needed quickly and easily.

Meaning: The exact definition of data fields relating to these tasks was easy to find out.

The exact meaning of data elements was either obvious, or easy to find out.

Right data: The system contained the data I needed to carry out these tasks.

Sufficiently detailed data was maintained by Sample Company to answer the questions.

Ease of use: It was easy to learn how to use the computer system to do these tasks.

The computer system and commands were convenient and easy to use.

Understandability: The arrangement or layout of the data on the several databases was hard to understand.

It was easy to figure out what data was available on a given subject.

Familiarity with a database system: Before this exercise, I already had some experience using DBase III.

Familiarity with programming: Before this exercise, I was familiar with at least one programming language.

Appendix C. Empirical results

C.1. H1: TTF—performance

One unexpected anomaly in the data had to be compensated for by adjustments in our analysis tools. Sixty percent of the subjects had accuracy scores of 100% in session 4, resulting in a non-normal, truncated distribution for the accuracy variable. This raises some concern about the appropriateness of regression as an analysis tool for accuracy, since it assumes that the expected value of all error terms is zero, which may not be the case with such a distribution. For this reason the analysis of accuracy in Table 5 shows the results for both regression and logistic regression [8], which is not sensitive to violations of this assumption. The logistics regression analysis uses maximum likelihood estimations to predict the odds of being in the 100% accurate group versus not being in the 100% accurate group, based on the same independent variables as used in the regression.

We also computed results when the treatment variables (technology fit with task, individual fit with technology) and control variables (pair tendency to work fast, pair tendency to work accurately, lab instructor, previous experience with a database language, previous experience with programming) are regressed first against time-to-complete, and then against accuracy. The third column shows the results of the logistic regression for accuracy.

The regression explaining time-to-complete was significant at the 0.0001 level, explaining 31% of the variance (with an adjusted $R^2$ of 0.25). On average, students with the integrated environment completed the exercise 5 min faster. Those who had training on the same technology in a previous session completed the exercise 7 min faster. Both effects were significant at the 0.001 level, supporting H1a and H1c. A separate analysis showed no interaction effect.
Two of the five control variables were significant in explaining time-to-complete—the pair tendency to work fast (significant at 0.01), and the lab instructor (significant at 0.01).

The regression explaining accuracy was also significant at the 0.0001 level, explaining 31% of the variance (with an adjusted $R^2$ of 0.25). The logistic regression analysis mirrors the results found by the regression analysis, with slightly greater significance for the impact of integration on accuracy. The $R^2_L$ statistic in logistic regression is a measure of predictive efficacy. It varies from 0 to 1, and is roughly analogous to the $R^2$ statistic in ordinary least squares regression, though it is not strictly interpretable as the percent of variance explained. However, the significance of the overall analysis, as indicated with asterisks next to the $R^2_L$ statistic, is analogous to the significance shown for the ordinary least squares [8].

Logistic regression estimates give the log odds of being in the 100% accurate group versus not being in the 100% accurate group. With an estimate of 0.70 for the coefficient of ‘integrated environment’, the impact on the log odds is the addition of 0.70, and the impact on the odds is the multiplication by 2.0 (i.e. $e^{0.70}$).

Thus, the subjects in the integrated environment were about twice as likely as non-integrated subjects to have 100% accuracy (based on the logistics regression analysis). Alternatively, based on the regression, subjects in the integrated environment were, on average, about 13% more accurate. The treatment of session 3 training did not seem to affect accuracy. Once again, a separate analysis showed no interaction effects. Of the control variables, only the pair tendency to work accurately is statistically significant, at the 0.0001 level.

Thus, H1b is supported, but H1d is not. These results allow us to be more specific about what we expect to find in analyzing the data for H3. Specifically, since the integrated data (higher TTF) improved both time-to-complete and accuracy, we would expect that user evaluations of the consistency of the data would also predict both time-to-complete and accuracy. We can label these more specific hypotheses H3a and H3b. Since training on the same system (higher TTF) affected time-to-complete but not accuracy, we would expect that user evaluations of the adequacy of training would predict time-to-complete, but not accuracy. This we can label H3c.

### Table 5
Regression results for H1: determinants of time-to-complete and accuracy

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Regression</th>
<th>Logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time-to-complete (min)</td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>N</td>
<td>114</td>
<td>127</td>
</tr>
<tr>
<td>$R^2$ ($R^2_L$ for logistic regression)</td>
<td>0.31****</td>
<td>0.31****</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Regression B Estimates</td>
<td>Logistic regression estimates</td>
<td></td>
</tr>
<tr>
<td>Treatment variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology fit with task (integrated data)</td>
<td>$-5.22^{***}$</td>
<td>0.13*</td>
</tr>
<tr>
<td>Individual fit with technology (training on same data environment)</td>
<td>$-7.35^{***}$</td>
<td>$-0.02$</td>
</tr>
<tr>
<td>Control variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pair tendency to work fast</td>
<td>2.31**</td>
<td>$-0.03$</td>
</tr>
<tr>
<td>Pair tendency to work accurately</td>
<td>$-0.06$</td>
<td>0.19****</td>
</tr>
<tr>
<td>Lab instructor</td>
<td>**</td>
<td>NS*</td>
</tr>
<tr>
<td>Experience on another database</td>
<td>$-0.06$</td>
<td>0.00</td>
</tr>
<tr>
<td>Programming experience</td>
<td>$-0.41$</td>
<td>$-0.02$</td>
</tr>
</tbody>
</table>

* Dummy variable group for lab instructor not significant.
* 0.05.
** 0.01.
*** 0.001.
**** 0.0001.
Table 6
Regression results for H2: determinants of UE of consistency and training

<table>
<thead>
<tr>
<th>Regression</th>
<th>UE of consistency</th>
<th>UE of training</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>118</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.30 ***</td>
<td>0.35 ****</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.24</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Regression B estimates

<table>
<thead>
<tr>
<th>Treatment variables</th>
<th>UE of consistency</th>
<th>UE of training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology fit with task (integrated data)</td>
<td>1.00 ****</td>
<td>−0.13</td>
</tr>
<tr>
<td>Individual fit with technology (training on same data environment)</td>
<td>0.27</td>
<td>1.11 ****</td>
</tr>
</tbody>
</table>

Control variables

<table>
<thead>
<tr>
<th>Control variables</th>
<th>UE of consistency</th>
<th>UE of training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair tendency to work fast</td>
<td>−0.06</td>
<td>−0.12</td>
</tr>
<tr>
<td>Pair tendency to work accurately</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>Lab instructor</td>
<td>NS a</td>
<td>**</td>
</tr>
<tr>
<td>Experience on another database</td>
<td>0.08</td>
<td>0.25 ****</td>
</tr>
<tr>
<td>Programming experience</td>
<td>0.10</td>
<td>0.12 *</td>
</tr>
</tbody>
</table>

Table 7
Regression results for H3: UE of TTF predicting time-to-complete and accuracy

<table>
<thead>
<tr>
<th>Regression</th>
<th>Logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>103</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.25 **</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Regression B estimates

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Time-to-complete (min)</th>
<th>Accuracy (%)</th>
<th>Accuracy (100% accurate vs. not 100% accurate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UE for consistency</td>
<td>−2.27 **</td>
<td>−0.01</td>
<td>0.22</td>
</tr>
<tr>
<td>UE for training</td>
<td>−1.29</td>
<td>−0.02</td>
<td>−0.15</td>
</tr>
</tbody>
</table>

Logistic regression estimates

<table>
<thead>
<tr>
<th>Control variables</th>
<th>Time-to-complete (min)</th>
<th>Accuracy (%)</th>
<th>Accuracy (100% accurate vs. not 100% accurate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair tendency to work fast</td>
<td>2.21 **</td>
<td>−0.04</td>
<td>−0.36</td>
</tr>
<tr>
<td>Pair tendency to work accurately</td>
<td>0.52</td>
<td>0.20 ****</td>
<td>1.03 ***</td>
</tr>
<tr>
<td>Lab instructor</td>
<td>NS a</td>
<td>NS a</td>
<td>NS a</td>
</tr>
<tr>
<td>Experience on another database</td>
<td>0.62</td>
<td>−0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Programming experience</td>
<td>0.19</td>
<td>−0.03</td>
<td>−0.06</td>
</tr>
</tbody>
</table>

Table 6 shows the impact of the treatment and control variables on the two key dimensions of user evaluations of TTF, consistency and training (arrow H2 in Fig. 1). The regression model explained 30 and 35% of the variance, respectively (with adjusted $R^2$ of 0.24 and 0.29). User evaluations of consistency of the

C.2. H2: TTF→user evaluations of TTF

evaluations of TTF, consistency and training (arrow H2 in Fig. 1). The regression model explained 30 and 35% of the variance, respectively (with adjusted $R^2$ of 0.24 and 0.29). User evaluations of consistency of the

Table 6 shows the impact of the treatment and control variables on the two key dimensions of user evaluations of TTF, consistency and training (arrow H2 in Fig. 1). The regression model explained 30 and 35% of the variance, respectively (with adjusted $R^2$ of 0.24 and 0.29). User evaluations of consistency of the
data sources were quite significantly predicted by whether individuals used the integrated or the non-integrated system. UE ratings of the sufficiency of training were significantly explained by whether subjects had experience in session 3 with the same technology they were asked to use in session 4. User ratings of training are also significantly influenced by three of the control variables — database experience, programming experience, and lab instructor. We view this as good support for H2a and H2b: both of the key UE constructs were significantly linked to their expected precursor treatments.

C.3. H3: User evaluations—performance

Finally, Table 7 shows the link between user ratings of TTF and objective performance. For time-to-complete, 25% of the variance was explained, with an adjusted $R^2$ of 0.17. User evaluations of consistency were a strong predictor of time-to-complete, significant at 0.01. Contrary to expectations, UE of training adequacy was not significantly linked to time-to-complete, though the regression coefficient was relatively large and in the expected direction. (Its significance was 0.12.) Of the control variables, only the pair tendency to work fast was a significant predictor. Thus, H3a is supported, but H3b is not.

For accuracy, the analysis used both regression and logistic regression, because of the truncated distribution of accuracy scores. Both approaches told the same story. The regression model explained 30% of the variance in accuracy with an adjusted $R^2$ of 0.24. Neither UE of consistency nor UE of training was a significant predictor of accuracy. (The statistical significance for the two terms were 0.30 and 0.56, respectively). For both analysis techniques, the pair tendency to work accurately was the only significant predictor. It was not expected that evaluations of training would predict accuracy, but it was expected that evaluations of consistency would. Thus, H3c is not supported.

References

[21] M.T. Iaffaldano, P.M. Muchinsky, Job satisfaction and job


Dale L. Goodhue is an Associate Professor of MIS at the Terry College of Business at the University of Georgia. He received his PhD in MIS from MIT, and has published in Management Science, MIS Quarterly, Decision Sciences, Sloan Management Review and other journals. He is an Associate Editor for Management Science and the MIS Quarterly. His research interests include the impact of information systems, the impact of task-technology fit on individual performance, and the management of data and other IS infrastructures/resources. Current projects include understanding the impact of Data Warehousing and ERP systems.

Barbara D. Klein is an Assistant Professor of MIS at the University of Michigan-Dearborn. She received her PhD in Information and Decision Sciences from the University of Minnesota and has published in MIS Quarterly, Omega, Computers and Industrial Engineering, the Information Resources Management Journal, Database, and other journals. Her research interests include information quality and user error behavior. She has worked in the information systems field at IBM, Exxon, and AMP.

Salvatore T. March is a Professor in the Information and Decision Sciences Department, Carlson School of Management, University of Minnesota. He received a BS in Industrial Engineering (1972) and MS and PhD degrees in Operations Research (1975, 1978) from Cornell University. His primary teaching responsibilities and research interests are in the areas of Information System Development, Logical and Physical Database Design, Distributed Database Design, and Object-Oriented Languages, Development Tools, and Methodologies. His research has appeared in journals, such as ACM Computing Surveys, ACM Transactions on Database Systems, Communications of the ACM, IEEE Transactions on Knowledge and Data Engineering, The Journal of MIS, Information and Management, and Information Systems Research. He has served as Associate Editor and as Editor-in-Chief for ACM Computing Surveys and as an Associate Editor for MIS Quarterly. He is currently serving as an Associate Editor for Information Systems Research, Decision Sciences Journal, and Information Systems Frontiers. He has been involved in organizing committees and program committees for numerous conferences including the Entity-Relationship Conference (ER), ACM SIGMOD Conference, IEEE Data Engineering Conference, the Workshop on Information Technologies and Systems (WITS), and the International Conference on Information Systems (ICIS). He is currently a member of the Steering Committee for WITS and a member of the Program Executive Committee for ICIS.