*Information Systems Research*
Vol. 20, No. 4, December 2009, pp. 527–546
ISSN 1047-7047 | EISSN 1526-5536 | 09 | 2004 | 0527

informs®

DOI 10.1287/isre.1080.0198


# How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations

Ujwal Kayande
College of Business and Economics, Australian National University, Canberra, ACT 0200, Australia,
ujwal.kayande@anu.edu.au

Arnaud De Bruyn
ESSEC Business School, 95000 Cergy-Pontoise, France, debruyn@essec.fr

Gary L. Lilien, Arvind Rangaswamy
Smeal College of Business, Pennsylvania State University, University Park, Pennsylvania 16802
{glilien@psu.edu, arvindr@psu.edu}

Gerrit H. van Bruggen
Rotterdam School of Management, Erasmus University, 3000 DR Rotterdam, The Netherlands,
gbruggen@rsm.nl

Model-based decision support systems (DSS) improve performance in many contexts that are data-rich, uncertain, and require repetitive decisions. But such DSS are often not designed to help users understand and internalize the underlying factors driving DSS recommendations. Users then feel uncertain about DSS recommendations, leading them to possibly avoid using the system. We argue that a DSS must be designed to induce an alignment of a decision maker's mental model with the decision model embedded in the DSS. Such an alignment requires effort from the decision maker *and* guidance from the DSS. We experimentally evaluate two DSS design characteristics that facilitate such alignment: (i) feedback on the *upside potential* for performance improvement and (ii) feedback on *corrective actions* to improve decisions. We show that, *in tandem*, these two types of DSS feedback induce decision makers to align their mental models with the decision model, a process we call deep learning, whereas individually these two types of feedback have little effect on deep learning. We also show that deep learning, in turn, improves user evaluations of the DSS. We discuss how our findings could lead to DSS design improvements and better returns on DSS investments.

*Key words*: decision support systems; DSS design; feedback; learning; mental models; evaluations
*History*: Ritu Agarwal, Senior Editor; Giri Kumar Tayi, Associate Editor. This paper was received on February 20, 2008, and was with the authors $8\frac{1}{2}$ months for 2 revisions. Published online in *Articles in Advance* December 18, 2008.

## 1. Introduction

Technological and modeling advances have dramatically increased the availability and quality of model-based decision support systems (DSS) (Shim et al. 2002, Banker and Kauffman 2004). Many such systems (e.g., customer relationship management systems, retail marketing mix DSS, employee scheduling DSS, clinical prescription DSS, etc.) are designed to assist decision makers in environments in which: (i) the data available to aid decision making are voluminous and beyond human information processing capabilities, (ii) the link between decisions and outcomes is probabilistic or uncertain, and (iii) the decisions are repetitive. In such environments, it is highly unlikely that decision makers can consistently outperform recommendations from even a simple model-based DSS (Hoch and Schkade 1996). Yet, Umanath and Vessey (1995, p. 796) observe that "since human decision makers do not know the rationale behind the suggested recommendation, they are typically skeptical of the output produced and are therefore reluctant to use such systems." We examine whether a DSS will be

perceived as more valuable if it enables users to internalize the rationale behind those recommendations. We use our findings to develop insights on how DSS should be designed to enable such internalization.

There are many well recognized examples of user resistance to (objectively good) model-based DSS in data-rich, uncertain environments. Managers in retail grocery chains must set prices daily for thousands of products, integrating information about retail price elasticities amid uncertain competitive reactions. Retail pricing DSS that include price-optimization models have been shown to dramatically outperform retail managers (Reda 2003, Montgomery 2005). Yet, Sullivan (2005) reports that only 5% to 6% of retailers use such DSS, with most managers preferring to use gut feeling for pricing decisions. Similarly, clinical DSS significantly improve clinical performance in prescribing decisions (Hunt et al. 1998), yet medical professionals are largely unwilling to use them (Sintchenko et al. 2004, Lai et al. 2006). Ashton (1991), Singh and Singh (1997), and Sieck and Arkes (2005) among others have noted decision makers' disinclination to use DSS in a variety of different environments, even when the models embedded in the systems are known to improve decision quality and performance. Several researchers have suggested that a lack of user understanding of the logic underlying DSS output leads to poor perceptions of the value such model-based DSS offer, leading to user resistance and impeded system use (e.g., McIntyre 1982, Davis 1989, Van Bruggen et al. 1996, Lilien et al. 2004). In the context of retail pricing DSS, Montgomery (2005, p. 375) suggests that "Unless the model can provide some intuition in understanding why this new strategy is better, users are more apt to reject it." Indeed, Sanders and Manrodt (2003) found that 83% of a sample of forecasting managers considered "easy understandable results" to be the most important forecasting software feature, while 66% reported dissatisfaction with the software they currently used.

We propose that decision makers will be more likely to accept a DSS when their mental models[1] of the decision environment become aligned with the decision model embedded in the DSS (hereafter referred to as the *DSS model*). The literature provides some support for this view. Gonul et al. (2006) show that confident and long explanations associated with DSS advice can improve user acceptance of that advice. In the context of medical diagnosis of acute cardiac ischemia, Lai et al. (2006) found that a tutorial on the advice given by a clinical DSS increased the use of that advice by emergency care physicians, leading to better patient outcomes. Limayem and DeSanctis (2000) find that system explanations improve group DSS usability, particularly because of improvements in user understanding of decision models.

For mental models to be aligned with the DSS model, decision makers need decisional guidance (Silver 1991). However, Todd and Benbasat (1999) argue that decision makers also have to be induced to exert effort to change decision strategies, which reflect their mental models of the decision environment. We show that a *dual-feedback* DSS, which incorporates feedback *both* about upside potential (i.e., how much more can be gained by internalizing the DSS model) *and* feedback on corrective actions (i.e., guidance on how the manager's mental model should be corrected), would induce more effort from decision makers as well as offer appropriate decision guidance. This combination of effort and guidance then produces significant mental model updating, while single feedback DSS produce little or no updating. Mental model updating, in turn, leads to better subjective DSS evaluations than when little or no mental model updating occurs. While many DSS incorporate some form of feedback, our results show that DSS evaluations only improve after significant mental model updating, which occurs when the DSS incorporates *both* upside potential and corrective feedback.

We proceed as follows. We first present a conceptual framework explaining why the gap between the user's mental model and the DSS model influences

[1] A mental model is an individual's cognitive representation of a domain that supports understanding, reasoning, and prediction (Gentner and Stevens 1983, Norman 1983). The mental model representation of the task is then based on the decision maker's previous experiences and current observations, which provide the framework for how that decision maker performs the task (Wilson and Rutherford 1989, Lim et al. 1997). While the concept of a mental model can consist of many different aspects (including how to use a DSS), we define a decision maker's mental model of a decision domain here narrowly, as a cognitive representation of how multiple decision variables affect performance outcomes.

DSS evaluation. Next we propose a model of how dual feedback on upside potential and corrective actions should influence the updating of users' mental models. We then develop and test specific hypotheses in a realistic, but controlled, experimental setting. We conclude by discussing our research contributions.

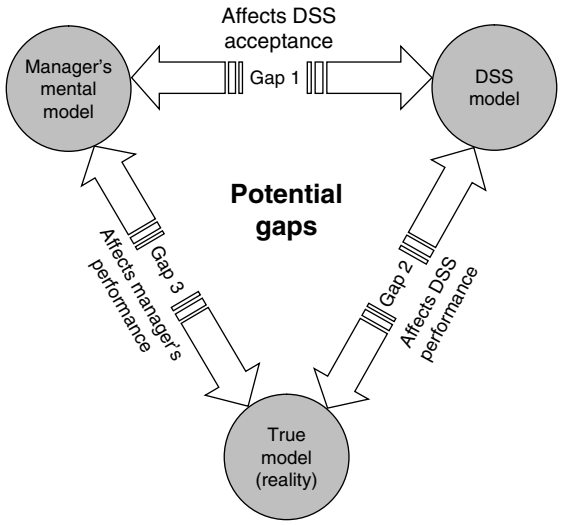
## 2. Mental Model Changes, DSS Evaluation, and DSS Design

### 2.1. The Effects of Mental Model Changes on DSS Evaluation

The 3-Gap framework (Figure 1) summarizes our perspective on the DSS evaluation problem.[2] Although we will use managerial decision making as the context, our framework is designed to apply to any data-rich domain where decisions are repetitive and outcomes are uncertain. We hypothesize that the magnitudes of the gaps between three models of the decision environment—the manager's mental model, the DSS model, and the unknown true model (which generates data in the real world, but is only partially observed ex-post)—determine the managers' decisions, the consequent outcomes, and DSS evaluations. To provide high-quality decision support, the gap between the DSS model and the true model must be small[3] (Gap 2 in Figure 1).

When users of a high quality DSS do not understand the rationale behind its recommendations, the gap between the DSS model and the user's mental model of the decision environment is likely to be large (Gap 1 in Figure 1). Consequently, the DSS model's recommended course of action and that implied by the user's mental model are likely to conflict, resulting in decision uncertainty (Einhorn and Hogarth 1980). Based on risk-adjusted preference theory (Keeney and Raiffa 1976), we propose that the objective quality of the DSS is then likely to be discounted by a risk-averse individual to account for the high uncertainty, leading to poorer subjective evaluations. Therefore, one potential source of the DSS evaluation problem lies in the inability of current DSS designs to close the gap between the user's mental model and the DSS model (Gap 1 in Figure 1). As a consequence, we suggest that the greater the change of the mental model in the direction of the DSS model, the better is the evaluation of the DSS that is used to effect the change (formalized later as H1). We focus on how to reduce Gap 1 because we hypothesize that this gap affects the user's evaluation of the DSS. We assume that the DSS model is of high objective quality (small Gap 2) and that it is of better quality than the user's mental model (large Gap 3). (We discuss this assumption in §4.2.)

**Figure 1** **The 3-Gap Framework: The Effect of Gaps Between Mental Model, DSS Model, and True Model**




### 2.2. Effects of Feedback on Mental Model Changes (Reducing Gap 1 in Figure 1)

We propose that to be recognized by users as valuable, thereby generating favorable evaluations, a DSS must be designed to incorporate characteristics that effect a change in the user's mental model, while improving his/her performance. The change in mental models could be of at least two types—(i) a relatively permanent deep change, or (ii) a transient change that disappears when the DSS is unavailable. We define these changes as follows:

*Deep learning* is a change in an individual's mental model that endures over time and/or over changes

[2] While gap analysis frameworks have been used in other contexts to understand, diagnose, and improve business and technology performance—see, for example, Parasuraman et al. (1985)—our framework focuses explicitly on the DSS evaluation problem.

[3] We recognize that the real world data generating process is not observable, so any DSS model, however good, is only a stylized representation of the process. The accuracy of a model, thus, might be best judged by how well it predicts the outcomes in a decision environment and/or how well the model fits past data.

in conditions—in other words, a change that concerns "the relatively permanent acquisition of skills, understanding, and knowledge" (Goodman 1998, p. 224).

*Shallow learning* is a change in an individual's mental model that occurs "only in the presence of external feedback or other conditions of practice, but disappears over time or when the supportive conditions are eliminated" (Goodman 1998, p. 224; also see Kluger and Denisi 1996, p. 278).

Deep learning will tangibly reduce the uncertainty about the DSS recommendations, i.e., reduce (cognitive) dissonance resulting in improved DSS evaluations, whereas shallow learning will not reduce DSS uncertainty. Our main interest in this paper is in how deep learning occurs because we expect it to affect DSS evaluations. Goodman et al. (2004) suggest that deep learning is most likely to occur when individuals are (i) motivated to, and actually exert *effort* to change their mental model, *and* (ii) provided *guidance* on how to modify that mental model, leading to deep learning. We formalize the joint effect of effort and guidance on deep learning as H2. Next we describe how each type of feedback individually influences effort and guidance, and therefore affects learning.

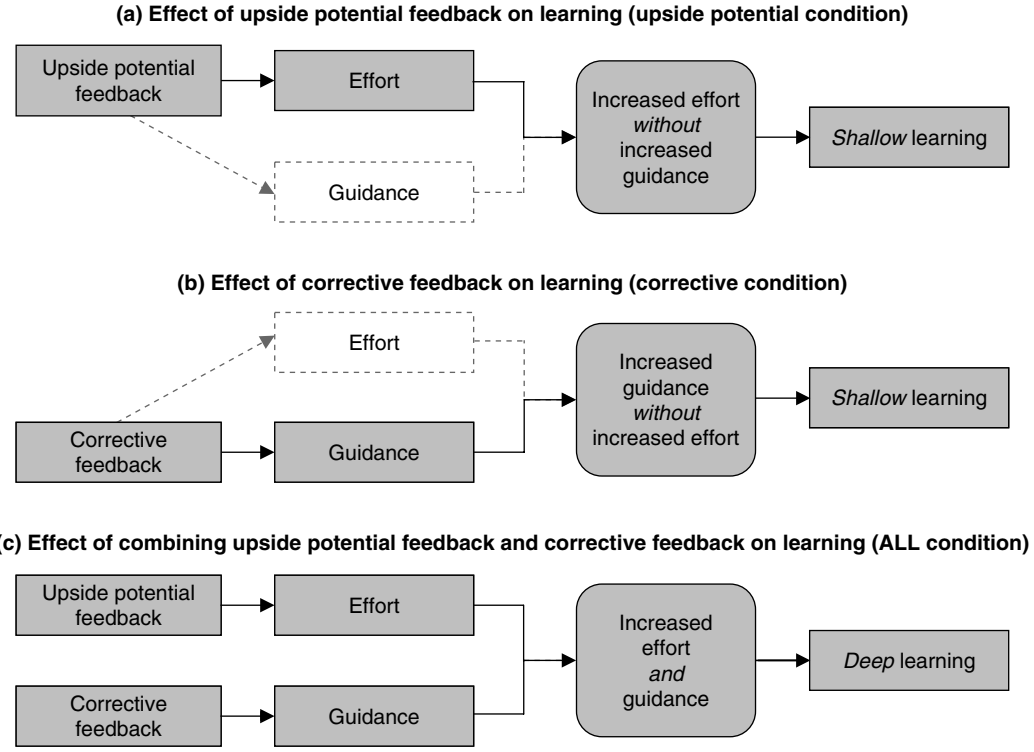
**2.2.1. Effects of Upside Potential Feedback on Effort and Learning.** Information about upside potential addresses how much better a manager might perform relative to current performance. For example, the sales module of the Siebel CRM system provides a salesperson with information on the sales achieved by the best performing salesperson and the sales levels in the best performing sales territory, providing proxies for upside potential. Chenoweth et al. (2004) show that users of decision support systems exert more effort to learn complex models when they know the upside potential. Upside potential feedback helps managers set specific (and challenging) goals, which drive increased effort to achieve them (Locke et al. 1981, Bandura 1997).

Several researchers have shown that while effort increases with more challenging goals, increased effort does not necessarily lead to deep learning because such goal-oriented behavior can focus the individuals' attention on the self, rather than on the task (Wood et al. 1990). As a result, task-learning processes are not activated (Kluger and Denisi 1996), leading to shallow learning and poor out-of-task performance. Upside potential feedback helps the manager set specific and challenging goals (e.g., match the best salesperson's performance), but does not provide the feedback necessary to learn *how* to perform better. Earley et al. (1990) found that the link between goal-setting, learning, and performance is greatly enhanced when individuals are provided with feedback about how to correct their strategies.

In summary, upside potential feedback will induce increased effort but may direct attention away from task-learning processes, resulting in *increased effort without appropriate learning*. So if upside potential feedback were to be combined with feedback that focuses attention on the task, we would expect managers to exert the increased effort and obtain the guidance necessary to obtain significant deep learning as summarized in Figure 2(a).

**2.2.2. Effects of Corrective Feedback on Guidance and Learning.** Corrective feedback, also called process feedback (Earley et al. 1990), can improve decision making, particularly in complex tasks, by increasing attention to task-learning processes and improving the quality of decision making (Balzer et al. 1992, Kluger and DeNisi 1996). This attention to task-learning processes improves performance. However, research also suggests that such feedback effects might only be transient—removal of such feedback can bring performance back to where it originally was (Goodman 1998, Goodman et al. 2004), meaning that DSS users will mechanistically implement DSS recommendations when they have a system available, but return to their traditional way of making decisions when the DSS is no longer available. Thus, corrective feedback might only lead to shallow learning because individuals directly adjust behavior by using the feedback rather than using the feedback to understand the task. For example, Atkins et al. (2002) find that if feedback is presented in a way that makes it trivial for decision makers to derive guidelines for action, they won't exert the effort needed to understand the rationale underlying these guidelines. Goodman et al. (2004) note that, "Essentially, feedback does the work for the performers, making it seemingly unnecessary

**Figure 2 Theoretical Framework Relating Feedback to Learning and Evaluation**




*Note.* Dotted lines indicate expectations of nonsignificant links.

for them to engage in the exploration, information-processing, and recall activities essential for learning." (p. 249).

Thus, corrective feedback directs attention to the task and task-learning process but also leads to less exploration and less effort. The result then is *increased guidance without increased effort*, resulting in low levels of deep learning, as summarized in Figure 2(b).

**2.2.3. Effects of Combining Upside Potential Feedback and Corrective Feedback.** Our arguments suggest that the two types of feedback should be viewed as *complementary* mechanisms; if the two feedback mechanisms are combined, the result should be an increase in guidance *and* effort, leading to deep learning (i.e., an alignment of the mental model towards the DSS model) as summarized in Figure 2(c).

### 2.3. Theoretical Model and Hypotheses

Figure 3 summarizes our theoretical framework, relating the two types of feedback to effort and guidance, deep learning, and DSS evaluation. It also provides a summary of the important empirical results discussed later. Our process model is as follows:

$$DSS\ Evaluation = \beta_{01} + \beta_{11} \cdot DeepLearning + \varepsilon_1 \quad \text{(M1)}$$

$$DeepLearning = \beta_{02} + \beta_{12} \cdot Effort + \beta_{22} \cdot Guidance + \beta_{32} \cdot (Effort \times Guidance) + \varepsilon_2 \quad \text{(M2)}$$

$$Effort = \beta_{03} + \beta_{13} \cdot \text{UPSIDE POTENTIAL FEEDBACK} + \beta_{23} \cdot \text{CORRECTIVE FEEDBACK} + \varepsilon_3 \quad \text{(M3)}$$

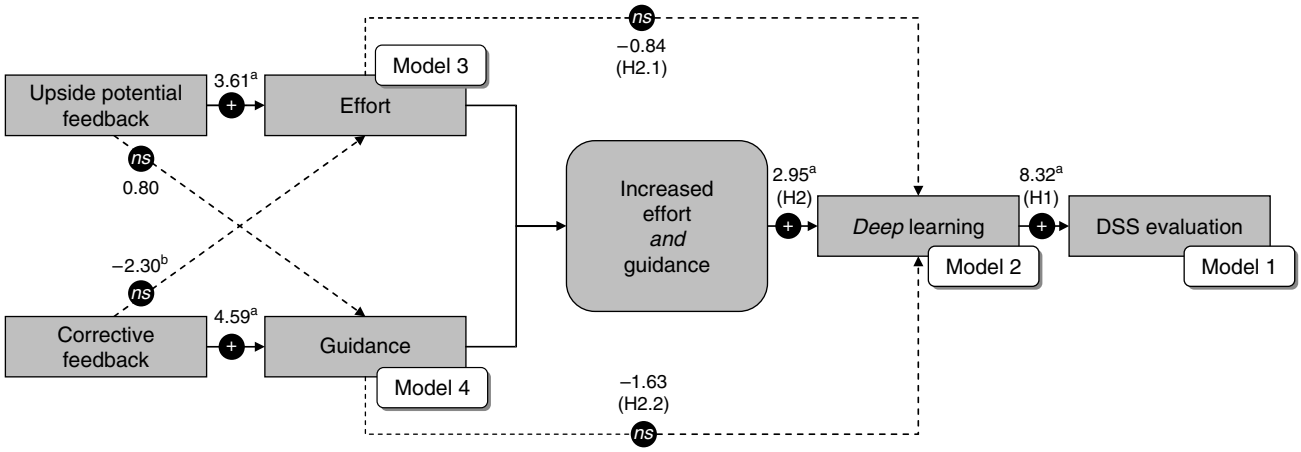
$$Guidance = \beta_{04} + \beta_{14} \cdot \text{UPSIDE POTENTIAL FEEDBACK} + \beta_{24} \cdot \text{CORRECTIVE FEEDBACK} + \varepsilon_4 \quad \text{(M4)}$$

Our first hypothesis relates DSS evaluation to the relatively permanent change in mental models:

HYPOTHESIS 1 (H1). *An increase in deep learning leads users to provide more favorable evaluations of the DSS. Therefore, we expect* $\beta_{11} > 0$.

We then hypothesize that it is the combination of increased effort and guidance that leads to deep learning.

**Figure 3 Connecting DSS Design Characteristics, Deep Learning, and DSS Evaluation (Models M1–M4)**




*Notes.* Dotted lines indicate expectations of nonsignificant links. *ns* stands for *no significant effect expected*; + stands for *positive effect expected*. We report *t*-statistics and statistical significance ([a]$p < 0.01$, [b]$p < 0.05$, two-tailed), as well as hypothesis numbering when appropriate.

Hypothesis 2 (H2). *The interaction of effort and guidance will have a positive effect on deep learning. Therefore, we expect* $\beta_{32} > 0$.

We also hypothesize that neither effort nor guidance alone leads to deep learning:

Hypothesis 2.1 (H2.1). *An increase in effort without guidance does not lead to deep learning. Therefore, we expect* $\beta_{12}$ *not to be statistically significant.* (*i.e.*, $\beta_{12} = 0$).

Hypothesis 2.2 (H2.2). *An increase in guidance without effort does not lead to deep learning. Therefore, we expect* $\beta_{22}$ *not to be statistically significant*, (*i.e.*, $\beta_{22} = 0$).

Per our discussion in §§2.2.1 and 2.2.2, effort increases when a manager is provided with feedback on upside potential, but effort is not expected to increase with corrective feedback, implying $\beta_{13} > 0$, but $\beta_{23} = 0$. On the other hand, guidance is influenced by the presence of corrective feedback but not by feedback on upside potential implying $\beta_{24}$ to be $>0$, but $\beta_{14} = 0$. Our expectations about these parameters serve as manipulation checks in our empirical analysis. Models M1–M4 comprise a test of the process model proposed in Figure 3.

## 3. Empirical Study

We sought to test our hypotheses using a realistic decision environment that is data rich, uncertain, and involves repetitive decisions by managers, criteria met by public charitable organizations soliciting donations via direct marketing (for example, World Vision or National Osteoporosis Foundation). Such organizations typically have access to very large data bases of past donors and prospects, see high uncertainty in response to any specific solicitation and conduct frequent, similar campaigns, leading to decisions that repeat both over time and across prospects.

We asked study participants to assume the role of a direct marketing manager of a large nonprofit charity focused on assisting people affected by natural disasters, and we provided them with a DSS to assist in their donor selection. Their main task was to identify the most attractive donors from a database of past donors for solicitation in a direct marketing campaign. Study participants were MBA students with direct marketing experience, as well as direct marketing managers working at charitable organizations similar to the one in our study.

We designed our empirical study to incorporate a challenging set of criteria. We sought:

(i) a decision environment that would have large data availability, unpredictability, and require repetitive decisions, so that managers would benefit from using a high-quality model-based DSS but not so complex as to be outside the skill range of our research participants;

(ii) a DSS whose underlying model sufficiently captures the real-world phenomenon (i.e., a small Gap 2 in Figure 1);

(iii) a context that would allow us to measure the user's mental model unobtrusively before, during, and after his or her interactions with the DSS;

(iv) a task in which we would be able to embed the DSS with each of the types of feedback (upside potential feedback and corrective feedback), both individually and jointly;

(v) a task that would allow us to measure deep and shallow learning unobtrusively; and

(vi) a task that would allow us to measure the process variables of interest (effort and guidance).

Criteria (i)–(iii) relate to the design of the overall context of the study, whereas criteria (iv)–(vi) relate to the design of the specific experiment to test our hypotheses. Our interest in testing how feedback affects the process of mental model changes (criterion iii) makes the design of a real-world field test challenging. Feedback must be accurate and immediate *across all experimental conditions*. This is difficult to obtain in the real world because of time delays between decisions and results, organizational and environmental noise, and lack of accurate information about what would have happened had other decisions been made (Tversky and Kahneman 1987). Therefore, to obtain both realism and control, we tested our hypotheses under controlled experimental conditions using a frequently occurring and realistic decision problem for which we could offer immediate feedback with known reliability and accuracy.

### 3.1. Experimental Context

Our experimental context was the solicitation of donations through direct mail for nonprofit or charitable organizations. In the United States alone, direct mail accounts for between $20 billion and $25 billion of the charitable educational and social change dollars contributed annually (Lister 2001). Direct marketing managers in charitable organizations typically solicit donations using large databases of potential donors. Each solicitation has a cost attached to it, and donation amount is donor-specific, so that it is critical for the direct marketing manager to identify the most likely (and high value) donors. This situation, in turn, requires the manager to understand the factors that influence the donor's likelihood of donation—that is, a mental model of the drivers of donation. DSS (e.g., MarketMiner Analyst™ website, www.modelingautomation.com) are often used by direct marketing managers to assist them in selecting high potential donors.

**3.1.1. Decision Environment.** To satisfy criterion (i), we sought a decision environment that would be sufficiently complex, but not outside the skill range of our participants. We designed a direct marketing decision environment complex enough to require the use of a DSS to select customers from a large database (200,000 in our case) of (hypothetical) donors, described on four characteristics—*recency* of donation (the number of quarters since their last donation), *frequency* of donation (the number of donations the donor has made in the past 5 years), *amount* of past donations (the average donation amount, in dollars, observed in the past for this particular donor), and the donor's *age*. The first three characteristics are often used by direct marketing firms in targeting models, typically referred to as Recency-Frequency-Monetary Value (RFM) model. We added age to the model to increase the complexity of the decision environment. Charities commonly use these factors to target donors (e.g., see Schlegelmilch et al. 1997). We modeled the probability, $p$, that a particular donor would make a donation, if solicited, by a logit function (Agresti 2002) as follows:

$$p = 1/(1 + \exp(5 - (X/20))), \tag{1}$$

where $X$ is called the donor's "attractiveness" and is given by

$$X = \beta_0 + (\beta_1 \times \textit{recency}) + (\beta_2 \times \textit{frequency}) + (\beta_3 \times \textit{amount}) + (\beta_4 \times \textit{age}). \tag{2}$$

The parameters of the "true" data generating model were $\beta = \{20, -20, 40, 10, 30\}$. We informed participants that donors were more likely to donate if they (1) had donated more recently, (2) had donated more frequently in the past five years, (3) had donated greater amounts, and (4) were older.

We generated a database of customers to satisfy two criteria. (1) The probabilities of donation in our database should be similar to those observed in actual not-for-profit databases. (2) The characteristics should be generated such that each donor could be described on a 0-to-100 attractiveness scale

with an average at the midpoint. The latter criterion ensured that we could subsequently ask participants to rate each donor on the same scale. To satisfy these two criteria, we generated donor characteristics from uniform distributions between 0 and 1 (after rescaling to account for differences in measurement units) independent of one another and incorporated these values within the functional form of the logit function described in Equations (1) and (2). As a result, a donor's true attractiveness varied between 0 and 100 with an average of 50, and donors' probability of donation varied between 0.67% and 50% with an average of 7.6%. Although average response rates, in practice, vary widely for different charities, an average response rate of 7.6% falls within industry averages for "warm" donors (see www.fundraising.co.uk/forum/thread.php?id=500).

Per criterion (ii), we sought a DSS model that would be close to the true data-generating model. Therefore, we designed Gap 2 to be small by constructing a DSS model that was identical to the true model in terms of weight of each factor. However, to ensure that actual donations could be predicted only approximately by the DSS model, we added a random noise term to the true model in Equation (1).

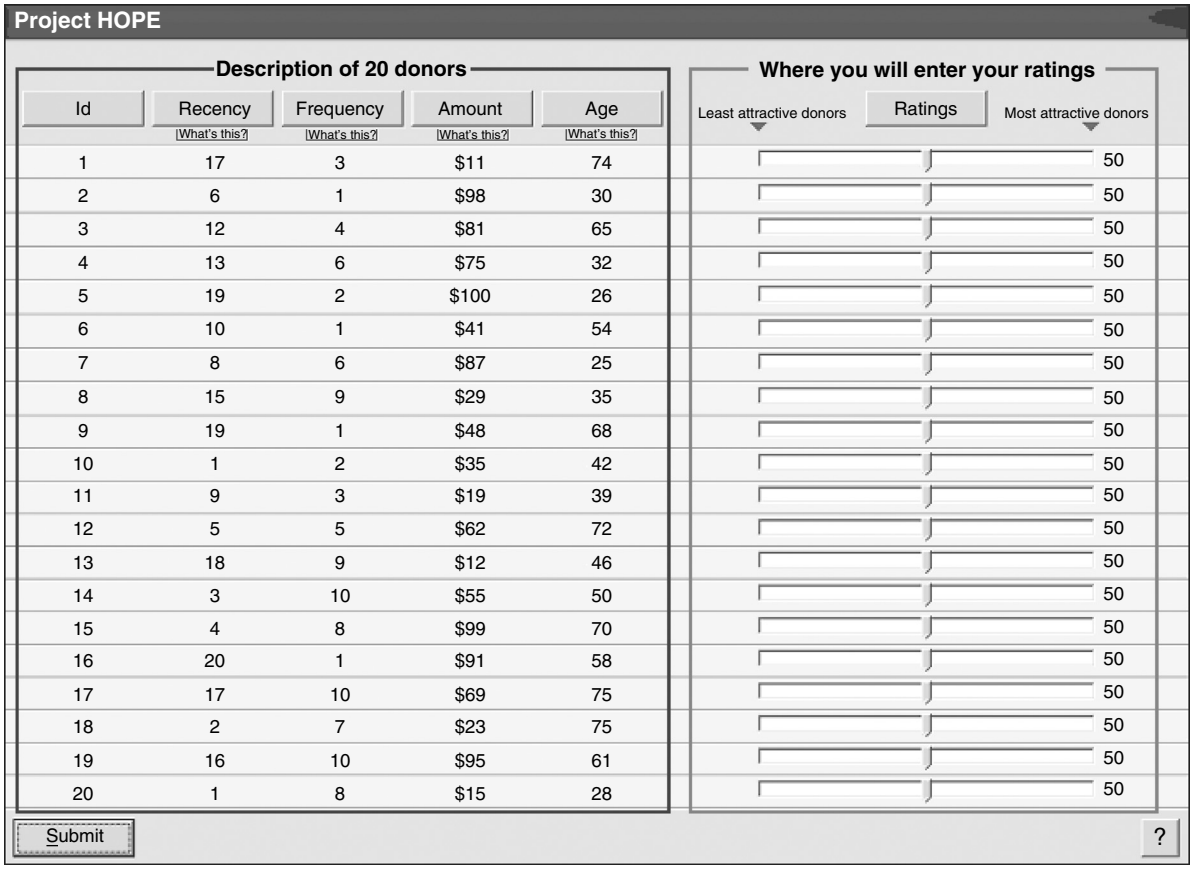
**3.1.2. Calibrating Participants' Mental Models.** To satisfy criterion (iii), we devised an unobtrusive and unbiased mechanism to calibrate each participant's mental model. We designed our study so that each participant made a sufficient number of decisions at each stage of study. This requirement allowed us to unobtrusively calibrate the participant's mental model, similar to Kunreuther's (1969) work on estimating managerial decision coefficients. Our unobtrusive approach minimizes potential biases compared to directly asking participants to reflect on their mental processes (Norman 1983). We asked each participant to rate 20 donors from the database on a 0 to 100 scale reflecting how attractive each donor was for selection in a marketing campaign. These 20 donors were described along the four drivers of donation behavior—recency, frequency, donation amount, and age (see Figure 4 for a screen shot of the task). This rating process (after rescaling and sorting) corresponds to the typical scoring mechanism that emerges in most direct marketing DSS (see D. Shephard Associates 1999).

To measure the participant's mental model, we statistically related their donor ratings to the descriptions of the 20 donors, thus inferring the implicit weights participants placed on the four factors.[4] Once a participant submitted his or her ratings, we estimated a linear regression model to determine the implicit weights $(\beta'_0, \beta'_1, \ldots, \beta'_4)$ that participant placed on recency, frequency, donation amount, and age. We then applied this calibrated mental model to the larger database of 200,000 donors to determine who to solicit. We told participants that each solicitation costs \$2 and, if successful, would generate a constant \$20 donation (to keep the task within participant skill range, criterion i), yielding a profitability threshold of 10% probability of donation. We applied the estimated mental model to the entire database, computing $X'$ and $p'$ for each of the 200,000 donors, and soliciting those donors with $p' > 0.1$. In addition to the marginal costs of solicitation, the fundraising campaign was subject to fixed costs of \$10,000. To determine whether a solicited donor actually makes a donation, we draw a random number $z$ from a uniform distribution $[0, 1]$ for each donor, and each solicited donor makes a donation of \$20 if $z$ is less than the true probability of donation ($p$). Note that if participants provided perfect scores ($X' = X$), mental model parameters would be equal to true parameters ($\beta' = \beta$), and the solicitation strategy would be optimal.

To assist participants in their decision making, we provided them with a DSS to select attractive donors from a database. We addressed the issue of incentive alignment by informing participants in all conditions that the amount of money they earned would be directly proportional to their financial performance. Participants were paid 0.015% of their financial performance on Task 1 and Task 2 (described more fully

[4] To estimate this relationship, there must be sufficient variation in the description of the 20 donors on each of the four factors and the factors must not be multicollinear, allowing for independent estimation of each weight. While a fractional factorial design is typically used in such cases, the number of profiles that our participants would have to rate made that approach infeasible. Therefore, we generated donors' characteristics (recency, frequency, etc.) so that extreme values were represented more often in the sample than in the population, while spanning the entire parameter space. To avoid multicollinearity, we randomly permuted donors' characteristics in the participant's rating sample, until no intercharacteristic correlation was higher than 0.15.

**Figure 4 DSS Interface, Illustrating the Respondent Task**




in the next section), in addition to $15 for participating in the study. The performance-based incentive was identical across all conditions.

### 3.2. Design of the Experiment (Manipulations and Measurements)

We summarize the sequence of steps in our experiment in Box A of Figure 5. Our experiment consisted of three main parts, addressing study design criteria (iv), (v), and (vi), respectively.

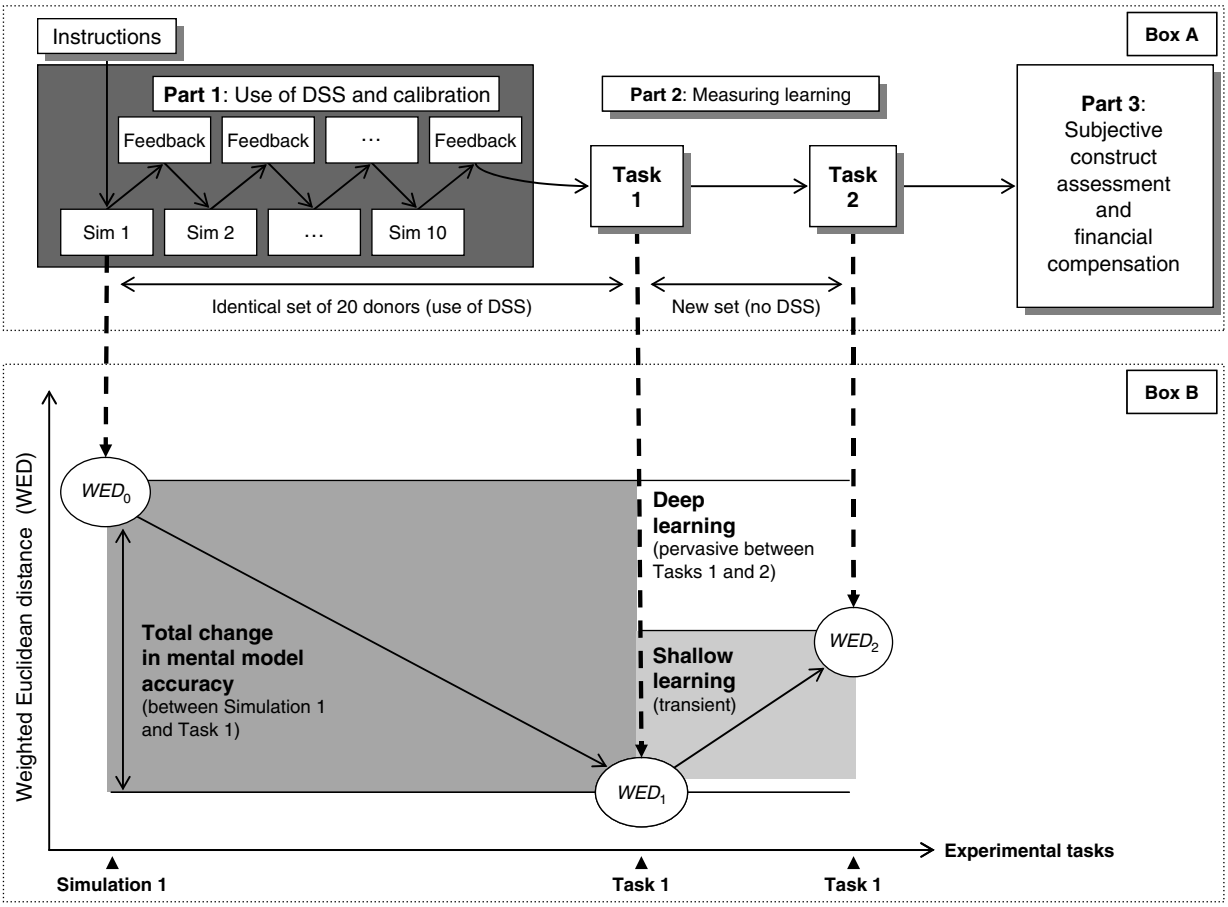
**3.2.1. Part 1: Using the DSS.** In Part 1 of the study, we asked participants to rate the same 20 donors in each of ten simulations. The participants had access to a DSS to help them determine the best possible ratings. In each simulation, participants rated the 20 donors, submitted those ratings to the DSS simulator, and obtained the DSS prediction of the performance of the campaign based on the participants' donor ratings. In the background, we calibrated the regression relationship between the participants' ratings and the description of the 20 donors on the 4 factors. The DSS was therefore both a support tool for users to make decisions and also a research tool to measure users' mental models.

We varied the feedback provided by the DSS to reflect the two types of feedback under study. We varied upside potential feedback at two levels (present or absent) and corrective feedback at two levels (present or absent), for a design with four cells. Both types of feedback were absent in the control condition. Our four conditions were:

1. "CONTROL CONDITION": The participant was only informed of the expected performance of the donor ratings. For example:

The DSS predicts that a marketing campaign based on your ratings would generate $76,654 in revenue.

**Figure 5 Experimental Sequence (Box A) and Measures of Learning (Box B)**




*Notes.* Dotted arrows indicate the mapping of tasks to measures. WED is the Weighted Euclidean Distance between the mental model and the true model. WED decreases as accuracy increases.

2. "UPSIDE POTENTIAL FEEDBACK": In addition to information about the expected performance, the participants in this condition were also informed of the maximum financial performance they could have achieved if they had been able to uncover the "true" attractiveness scores of the 20 donors. For example:

The DSS predicts that a marketing campaign based on your ratings would generate $76,654 in revenue.

The DSS predicts that it would be possible to generate up to $99,934 in revenue from this database.

3. "CORRECTIVE FEEDBACK": In addition to information about the expected performance of the donor rating strategy, participants in this condition were given feedback on whether they were placing too much or too little weight on each of the four factors. To operationalize this feedback, we compared the participants' mental model parameters to the parameters of the DSS model. For example:

The DSS predicts that a marketing campaign based on your ratings would generate $76,654 in revenue.

Here is some feedback that will help you improve your ratings. In developing your ratings for these donors:

—You assume a relationship between recency and donating behavior that is opposite to what is known.

—You are greatly overestimating the importance of frequency.

—You are underestimating the importance of age.

4. "ALL": In this condition, we provided participants with feedback on expected outcome, upside

potential, and corrective actions (conditions 1–3 above), in that order. For example:

The DSS predicts that a marketing campaign based on your ratings would generate $76,654 in revenues.

The DSS predicts that it would be possible to generate up to $99,934 in revenue from this database.

Here is some feedback that will help you improve your ratings. In developing your ratings for these donors:

—You assume a relationship between recency and donating behavior that is opposite to what is known.

—You are greatly overestimating the importance of frequency.

—You are underestimating the importance of age.

(Note that all feedback summaries were customized and dynamically generated based on the actual ratings provided by each participant.)

Participants were randomly assigned to one of the four feedback conditions. To provide an incentive for participants to focus on the task during the simulations, we informed them that they would be required, *after* completing the 10 simulations, to rate the same donors for a real direct mail campaign that we refer to as Task 1. We calibrated the participant's mental model each time he or she rated the set of 20 donors (i.e., in each of the 10 simulations, Task 1, and Task 2, described next).

**3.2.2. Part 2: Measuring Learning.** Our goal in Part 2 is to measure deep and shallow learning. Per our definition of deep learning, we sought a measure of mental model change that survives when feedback is removed. Therefore, in Task 2, we asked participants to rate 20 donors *who were different from those in Part* 1. To ensure they applied their (updated) mental model of donor behavior to this task, we told them that the 20 new donors were from the same database used in Task 1, so the extent to which each factor impacted donor behavior was the same for these new donors as it was for the donors in Task 1. Because we were only interested in measuring their final mental model at this stage, we did not provide the participants with access to the feedback and simulations component of the DSS.

We then constructed a measure of mental model accuracy—the distance between the true model (which in our study is identical, on average, to the DSS model) and the mental model. We sought a measure of mental model accuracy that reflects the participant's ability to judge the relative importance of those factors. An individual-level measure that satisfies this criterion is as follows:

$$WED_t = \left[\sum_j \omega_j \cdot (\beta'_{jt} - \beta_j)^2\right]^{0.5}, \tag{3}$$

where $t$ is the task ($t = 1, 2$), $WED_t$ is the Weighted Euclidean Distance between the mental model and the true model in task $t$, $\omega_j$ is the importance of the $j$th ($j = 1$–4 in our study) driver of donation behavior in the DSS model, $\beta'_{jt}$ is the mental model parameter associated with the $j$th driver of donation behavior in task $t$, and $\beta_j$ is the true parameter associated with the $j$th driver of donation behavior (in our study, $\omega_j = \beta_j$ on average). $WED_t$ is a measure of mental model accuracy (Gap 3) in our study (we note here that accuracy *increases* as $WED_t$ decreases). Weighting the distance between coefficients by $\omega_j$ implies that a mental model that is close to the true model on the most important drivers is better than a mental model that is close to the true model on the less important drivers. Note that per our study design, if the DSS model were to converge to the true model, then, on average, Gap 3 is identical to Gap 1 and $\omega_j = \beta_j$. (We obtained substantially equivalent empirical results with an equally weighted, simple Euclidean distance measure, suggesting that our results are not sensitive to our choice of metric.)

We provide a graphical explanation of our learning measures in Box B of Figure 5, relating those measures to each step of the experiment. The participant's initial mental model accuracy is measured by $WED_0$, calibrated using the mental model parameters from the first simulation in the same manner as $WED_t$. The difference between $WED_0$ and $WED_1$ is a measure of the change in mental model accuracy that is attributable to the participant's use of the DSS. Part of this change is a result of an internalization of the DSS model, i.e., deep learning, and part is a transient change that will disappear with the removal of the feedback, i.e., shallow learning. Asking participants to complete Task 2 gives us the ability to independently calibrate each part, explained next.

A rather persistent change in the mental model would be reflected in the extent to which the mental

model in Task 2 is more accurate than that in the initial simulation. Therefore, we construct a measure of deep learning by taking the difference between $WED_2$, mental model accuracy in Task 2, and $WED_0$, the accuracy of the initial mental model. We define deep learning (DL) as

$$DL = (WED_0 - WED_2). \quad (4)$$

In contrast to deep learning in Equation (4), if mental model accuracy in Task 1 is much greater than that in Task 2, it indicates that the accuracy of the mental model in Task 1 was a result of the participants being mechanistic in their approach to the task—an approach that would lead to decision quality deterioration if conditions were changed, as in Task 2, with a new set of donors. We define shallow learning (SL) as

$$SL = (WED_2 - WED_1). \quad (5)$$

While Parts 1 and 2 of our study allowed us to manipulate feedback provided by the DSS and measure participants' mental models, we also must measure process variables and DSS evaluations.

**3.2.3. Part 3: Subjective Construct Assessment.** The main process variables of interest are effort and guidance. After participants completed Tasks 1 and 2, but before they were informed of the financial performance results, we asked them to complete a questionnaire measuring effort (four items adopted from Lilien et al. 2004: "I was totally immersed in addressing this problem," "I took this task seriously," "I put in a lot of effort," and "I wanted to do as good a job as possible no matter how much effort it took"), and guidance ("The DSS gave clear guidance on how I could do better"). As a proxy measure of effort, we also recorded the time spent by each participant on the simulations and the two tasks. After being informed of their results, participants evaluated the DSS ("I would definitely recommend a DSS like the one I had available to direct marketers"). All items were 5-point Likert scale questions, with 1 = completely disagree and 5 = completely agree. In addition, participants were asked to respond to multiitem scales on perceived usefulness of the DSS, perceived ease of use of the DSS, perceived enjoyment of the task, decision confidence (two items: "I am in full agreement with the ratings I gave" and "I am confident that the ratings I gave will work out well"), and decision style. We also asked participants open ended questions about what they thought were the drivers of donation behavior, as well as their donor selection approach.

### 3.3. Sample and Experimental Procedure

We sought participants who would be appropriate surrogates for direct marketing managers. Such participants had to have *well-formed* mental models of the drivers of donor attractiveness prior to the experimental study. Hence, we recruited 81 MBA students at a large northeastern U.S. university, who had been exposed to RFM models in at least one of their courses, to assume the role of the direct marketing manager. We randomly assigned participants to one of the four conditions. We sought a measure to determine whether our participants' mental models were well formed (i.e., not random). One such measure is the $R^2$ associated with the calibration of each participant's initial mental model (simulation 1). The $R^2$ measure captures the consistency of the model used by the participant to rate each donor, indicating whether the initial mental model was well-formed (high $R^2$) or poorly formed (low $R^2$). The $R^2$ for our participants ranged from 0.13 to 0.99, with a trimodal distribution. A group of 3 participants had an $R^2$ between 0.13 and 0.20, a second group (6 participants) between 0.34 and 0.49, and a final group (72 participants) between 0.58 and 0.99. Based on the structure of the empirical distribution, we classified those participants with $R^2$ under 0.58 as having ill formed mental models and disqualified them, leaving us with data on 72 (primary) participants for analysis.

The average age of our participants was 27. Each participant had at least four years of work experience, and had been exposed to RFM models in one of their courses. Remus (1996) finds that such graduate students are good surrogates for managers, while undergraduate students with no experience are poor surrogates. However, others argue that graduate students are not always good surrogates for managers (Hughes and Gibson 1991). We used the approach of Nicolaou and McKnight (2006) and tested whether data from a sample of managers are consistent with the data from our student subjects. We recruited

10 direct marketing managers currently employed with charity organizations to participate in our study. We included a dummy variable in all our models (M1–M4) to test whether there was a mean difference between managers and students on the four dependent variables and found none. We also tested whether the pattern of data provided by each sample group on the four main variables of interest—evaluation, deep learning, effort, and guidance—was equivalent by testing for the equivalence of the data covariance matrices across the two sample groups. Bartlett's test (Anderson 1984; $\chi^2 = 19.83$ with 10 df, $p = 0.03$), as well as Box's $M$ Test (Box's $M = 19.18$; $F = 1.59$, ns), suggest equivalence of the two covariance matrices, indicating that our sample of students serve as appropriate surrogates for managers. The data equivalence also permitted us to pool the student and manager data for all our analyses. The final sample of 82 participants resulted in 16 to 23 participants per condition. Participants earned between \$23 and \$46, with the average payment being \$37. They took an average of 32 minutes to complete the two tasks, with a range of 13 to 77 minutes.

### 3.4. Analysis

Figure 3 and Models (M1)–(M4) sketch the process model of the effect of DSS design characteristics on deep learning and DSS evaluation. There are several submodels embedded in this framework, with errors that are likely to be correlated. For the empirical analysis, we included two additional terms in Model M1 to control for shallow learning and participants' individual financial compensation, as follows:

$$DSS\ Evaluation = \beta_{01} + \beta_{11} \cdot DeepLearning + \beta_{21} \cdot ShallowLearning + \beta_{31} \cdot Financial\ compensation + \varepsilon_1. \quad (6)$$

We estimated the parameters of the four models simultaneously using a full information maximum likelihood (FIML) routine in SAS, assuming correlated errors across all models.

### 3.5. Results

Results of estimating models (M1)–(M4) are shown in Table 1. (As noted previously, none of the manager-student dummy variable was significant at the 0.10 level; hence, for simplicity we exclude reporting them in our tables.) We hypothesized that users' evaluations of the DSS depend on the extent to which they had internalized the DSS model (H1; Part A of Table 1). We find strong support for H1 ($\beta_{11} = 0.060$, $p < 0.01$). We also find that shallow learning is not a significant driver of DSS evaluation ($\beta_{21} = -0.000$, *ns*), further supporting our theory that DSS evaluation depends on a *significant* updating of mental models, i.e., on deep learning. We also find that financial compensation is not a significant driver of evaluation ($\beta_{31} = 0.01$, *ns*). These results show that deep learning is a significant driver of DSS evaluation, after controlling for shallow learning and financial compensation, thus supporting H1.

Next, we tested H2, i.e., whether deep learning is affected by the combination of effort and guidance.

**Table 1 Relationship Between DSS Evaluation, Deep Learning, Effort, and Guidance**

| Variable | Coefficient | Beta | $t$-stat | Hypothesis |
|---|---|---|---|---|
| Part A: Effect of deep learning on DSS evaluation (Model M1) | | | | |
| Intercept | $\beta_{01}$ | 3.233 | 4.10[a] | |
| Deep learning | $\beta_{11}$ | 0.060 | 8.32[a] | H1 |
| Shallow learning | $\beta_{21}$ | 0.000 | −0.04 | |
| Financial compensation | $\beta_{31}$ | 0.010 | 0.49 | |
| Part B: Effect of effort and guidance on deep learning (Model M2) | | | | |
| Intercept | $\beta_{02}$ | −2.76 | −0.15 | |
| Effort | $\beta_{12}$ | −3.62 | −0.84 | H2.1 |
| Guidance | $\beta_{22}$ | −8.15 | −1.63 | H2.2 |
| Effort × guidance | $\beta_{32}$ | 3.50 | 2.95[a] | H2 |
| Part C: Effect of feedback type on effort and guidance | | | | |
| 1. Effort (Model M3) | | | | |
| Intercept | $\beta_{03}$ | 4.20 | 40.14[a] | |
| UPSIDE POTENTIAL FEEDBACK | $\beta_{13}$ | 0.38 | 3.61[a] | |
| CORRECTIVE FEEDBACK | $\beta_{23}$ | −0.25 | −2.30[b] | |
| 2. Guidance (Model M4) | | | | |
| Intercept | $\beta_{04}$ | 2.95 | 15.6[a] | |
| UPSIDE POTENTIAL FEEDBACK | $\beta_{14}$ | 0.16 | 0.80 | |
| CORRECTIVE FEEDBACK | $\beta_{24}$ | 0.91 | 4.59[a] | |

*Notes.* 1. Significant at: [a]$p < 0.01$, [b]$p < 0.05$ (two-tailed).

2. Deep learning is the most significant driver of DSS evaluation; supports H1.

3. The interaction of effort and guidance significantly affects deep learning; supports H2.

4. Effort increases significantly when upside potential feedback is provided, supporting the manipulation of effort with upside potential feedback.

5. Effort decreases significantly when corrective feedback is provided.

6. Guidance increases with corrective feedback but not with upside potential feedback, supporting the manipulation of guidance with corrective feedback.

The result (Part B of Table 1), supports H2 ($\beta_{32} = 3.50$, $p < 0.01$).[5] This result shows that the combination of effort and guidance is a significant driver of deep learning. The intercept term is not significant, indicating that deep learning does not occur without effort and guidance. The coefficients for effort and guidance are not significant, indicating that neither of these process variables alone is capable of obtaining deep learning. These results support H2.1 and H2.2.

Part C of Table 1 shows how the process variables were affected by the two feedback mechanisms (i.e., manipulation checks). Effort was significantly increased by the presence of upside potential feedback ($\beta_{13} = 0.38$, $p < 0.01$). In contrast, effort decreased with corrective feedback ($\beta_{24} = -0.25$, $p < 0.05$). Although we did not expect this significant negative effect, it is not entirely surprising considering the findings in the literature that corrective feedback leads to less inclination to exert effort (Atkins et al. 2002). Guidance, the other process variable, significantly increased in the presence of corrective feedback ($\beta_{24} = 0.91$, $p < 0.01$). As expected, guidance was not affected by the presence of upside potential feedback ($\beta_{14} = 0.16$, *ns*). (We also estimated Models (M1)–(M4) using "time taken to complete the simulations and tasks" as a proxy measure of effort, with similar results.)

Our Model (M2) results show that the interaction of effort and guidance produces deep learning: effort and guidance combine in complementary ways to help managers update their mental models. In Table 2, we show that the combination (i.e., Effort × Guidance) was significantly greater in the ALL condition than in the UPSIDE POTENTIAL FEEDBACK (*difference* = 3.22, $p < 0.01$, two-tailed) or CORRECTIVE FEEDBACK (*difference* = 2.01, $p < 0.05$, two-tailed) conditions. These aggregate results are consistent with our process model (see Figure 3).

Table 3 (Column A) summarizes the tests of whether there was significant deep learning in each of the four conditions. We find that deep learning is significantly different from 0 in the ALL condition (mean = 12.37, $p < 0.01$, two-tailed). There is evidence that deep learning in the UPSIDE POTENTIAL FEEDBACK condition (mean = 5.15, $p < 0.05$, two-tailed) and the CORRECTIVE FEEDBACK condition (mean = 4.62, $p < 0.05$, two-tailed) are both significantly different from 0. However, deep learning in both these conditions is significantly less than that in the ALL condition, consistent with our hypotheses (results shown under Column B of Table 3).

[5] In Model (M2), we also controlled for the time taken by the participants and obtained substantively similar results.

**Table 2 Effect of Upside Potential Feedback and Corrective Feedback on the Combination of Effort and Guidance**[1]

| Condition | Mean | Difference from ALL condition | *t*-stat |
|---|---|---|---|
| CONTROL condition | 12.67[2] | −4.87 | −3.67[a] |
| UPSIDE POTENTIAL FEEDBACK condition | 14.32[2] | −3.22 | −3.48[a] |
| CORRECTIVE FEEDBACK condition | 15.53[3] | −2.01 | −2.22[b] |
| ALL condition | 17.54 | | |

*Notes.* 1. We test here whether (Effort × Guidance) is greater in the ALL condition than in each of the other conditions.

2. Mean in this condition is significantly less than that for the ALL condition at [a]$p < 0.01$ (two-tailed).

3. Mean in this condition is significantly less than that for the ALL condition at [b]$p < 0.05$ (two-tailed).

**Table 3 Effect of Upside Potential Feedback and Corrective Feedback on Deep Learning**[1]

| | | A. Is deep learning significantly different from 0? | | B. Is deep learning significantly less than that in ALL condition? | |
|---|---|---|---|---|---|
| Condition | *N* | Average deep learning | *t*-stat | Difference | *t*-stat |
| CONTROL condition | 16 | 0.25 | 0.09 | −12.12 | −3.17[a] |
| UPSIDE POTENTIAL FEEDBACK condition | 23 | 4.62 | 2.01[b] | −7.76 | −2.62[b] |
| CORRECTIVE FEEDBACK condition | 21 | 5.15 | 2.17[b] | −7.22 | −2.40[b] |
| ALL condition | 22 | 12.37 | 5.17[a] | | |

*Notes.* 1. We test here whether (A) deep learning is significantly different from 0 in each of the conditions, and (B) deep learning is significantly less in each condition than that in the ALL condition.

2. Deep learning in UPSIDE POTENTIAL FEEDBACK condition is significantly different from 0 at [b]$p < 0.05$, and is significantly less than that in the ALL condition at: [b]$p < 0.05$ (two-tailed tests).

3. Deep learning in CORRECTIVE FEEDBACK condition is significantly different from 0 at [b]$p < 0.05$, and is significantly less than that in the ALL condition at: [b]$p < 0.05$ (two-tailed tests).

4. Deep learning in ALL condition is significantly different from 0 at: [a]$p < 0.01$ (two-tailed).

5. Indicates that deep learning in ALL condition is significantly greater than that in other conditions.

**Table 4 Effect of Upside Potential Feedback and Corrective Feedback on Shallow Learning**[1]

| | | A. Is shallow learning significantly different from 0? | | B. Is shallow learning significantly more than that in ALL condition? | |
|---|---|---|---|---|---|
| Condition | *N* | Average shallow learning | *t*-stat | Difference | *t*-stat |
| CONTROL condition | 16 | 4.52 | 3.27[a] | 1.61 | 1.39 |
| UPSIDE POTENTIAL FEEDBACK condition | 23 | 5.32 | 5.23[a] | 2.41 | 2.36[b] |
| CORRECTIVE FEEDBACK condition | 21 | 5.41 | 5.21[a] | 2.50 | 2.44[b] |
| ALL condition | 22 | 2.91 | 2.77[a] | | |

*Notes.* 1. We test here whether (A) shallow learning is significantly different from 0 in each of the conditions, and (B) shallow learning is significantly more than that in the ALL condition.

2. Shallow learning in UPSIDE POTENTIAL FEEDBACK condition is significantly different from 0 at [a]$p < 0.01$, and is significantly more than that in the ALL condition at: [b]$p < 0.05$ (two-tailed tests).

3. Shallow learning in CORRECTIVE FEEDBACK condition is significantly different from 0 at [a]$p < 0.01$, and is significantly more than that in the ALL condition at: [b]$p < 0.05$ (two-tailed tests).

Table 4 presents the analysis of whether significant shallow learning occurred in each of the conditions. We find that a significant level of shallow learning occurred in all the feedback conditions but that it was significantly greater in the UPSIDE POTENTIAL FEEDBACK condition (mean = 5.32) and the CORRECTIVE FEEDBACK condition (mean = 5.41) than in the ALL condition (mean = 2.91). The presence of shallow learning and deep learning in each of the conditions indicates that the observed mental model accuracy in Task 1 is partly based on a mechanistic approach specific to the 20 donors in the simulations and partly based on a real change in their mental model.

In Table 5, we report the financial performance and compensation of participants in each condition. Participants in the ALL condition performed significantly better than those in the other conditions. This result shows that participants provided with both types of feedback are more likely to perform better as well.

**Table 5 Effect of Upside Potential Feedback and Corrective Feedback on Financial Performance and Compensation**[1]

| Condition | Average financial performance in Task 1 ($) | Average financial performance in Task 2 ($) | Average financial compensation | Difference from ALL condition | *t*-stat |
|---|---|---|---|---|---|
| CONTROL | 76,328 | 61,583 | 35.98[2] | −3.66 | −2.79[a] |
| UPSIDE POTENTIAL FEEDBACK | 79,160 | 60,311 | 35.88[2] | −3.71 | −3.55[a] |
| CORRECTIVE FEEDBACK | 83,602 | 64,919 | 36.62[2] | −3.01 | −2.81[a] |
| ALL | 86,676 | 76,993 | 39.64 | | |

*Notes.* 1. We test here whether financial compensation of participants is greater in the ALL condition than in each of the other conditions.

2. Mean compensation in this condition is significantly less that for the ALL condition at [a]$p < 0.01$ (two-tailed).

3. We note that financial compensation is directly proportional to financial performance (participants were paid 0.015% of financial performance, plus $15 participation fee). The maximum financial performance possible in each task is about $100,000. We also note that we report here the average of actual compensation paid, which was not precisely 0.015% of financial performance because of rounding.

## 3.6. Validation Checks

**3.6.1. Deep Learning Effect on Confidence.** Per §2.1, our theoretical framework argues that users feel more confident when their mental models are updated towards the decision model, and it is this confidence (or lack of perceived uncertainty) that improves DSS evaluations. To see if indeed confidence was affected by deep learning, we included an additional model (M1A) in our model system, as follows:

$$DSSEvaluation = \beta_{01} + \beta_{11} \cdot Confidence + \beta_{21} \cdot FinancialCompensation + \varepsilon_1 \quad \text{(M1)}$$

$$Confidence = \beta_{01A} + \beta_{11A} \cdot DeepLearning + \beta_{21A} \cdot ShallowLearning + \varepsilon_{1A}. \quad \text{(M1A)}$$

We find that confidence is significantly affected by deep learning ($\beta_{11A} = 0.04$; $p < 0.01$) but not by shallow learning. In addition, we find that confidence has a significant positive effect on DSS evaluation ($\beta_{11} = 1.41$; $p < 0.01$), providing support to our theoretical argument. Participants in the ALL condition also had significantly greater confidence (mean = 4.30; $p < 0.01$) in their decisions than those in all other conditions.

**3.6.2. Mental Model Calibration Validity.** Our empirical approach is unique in the way we unobtrusively calibrated participants' mental models. To investigate the validity of our calibration, we compared the calibrated mental model in Task 2 with the participants' responses to an open ended question

asking them to list what they thought were the important drivers of donation behavior. We counted the number of times that the factor we calibrated in Task 2 as the most important driver was the one the participant mentioned as the most, or the second most, important driver. This count was 79%. We also counted the number of times that the factor we calibrated as least important in Task 2 was the one the participant mentioned as the least, or second least, important driver. This count was 63%. We then tested the extent to which our calibration's rank-ordering of factors was correlated with the participants' verbalization of the rank-ordering. The average Spearman correlation was 0.43, much greater than by chance (which would be close to 0). These results suggest that our mental model calibration appears to be an appropriate one.

**3.6.3. Alternative Model Specifications.** The model system in Figure 3 is a process model, so we sought to test it against two alternatives: a model specifying both direct and indirect effects of feedback on deep learning and an alternate model specification without process variables (effort and guidance). To assess the first alternative, we compared the fit of our model system (AIC = 1,220.48) with that of a system with additional direct effects of the two types of feedback on deep learning (AIC = 1,224.38). To assess the second alternative, we compared the fit of our model system (AIC = 1,220.48) against an alternate model system without process variables (AIC = 1,250.04). Our proposed model passed both these validity tests. Finally, we note that our results are also robust to the inclusion of variables such as perceived usefulness and perceived ease of use in Model (M1), indicating that the effects of feedback and deep learning on DSS evaluation are above and beyond the effect of variables previously studied in the literature.[6]

**3.6.4. Appropriateness of Subjects.** As discussed previously, our supplemental sample of domain-specific managers showed no significant differences, either in effect level or in covariance matrices for the four main variables, suggesting that the subject pool poses no major threats to the validity of our findings.

[6] Full details are available from the authors.

# 4. Discussion, Future Research, and DSS Design Implications

## 4.1. Discussion and Contributions

Researchers interested in the design and use of DSS have explored factors that lead to greater system usage and/or better performance. These factors include "task-technology" fit (Lim and Benbasat 2000), tabular versus graphical presentation format and color (e.g., Benbasat and Dexter 1985), fit between cognitive style and presentation format (Ramaprasad 1987), fit between information format and task (Vessey and Galletta 1991), accessibility (e.g., Mawhinney and Lederer 1990), adaptability/flexibility (e.g., Udo and Davis 1992), perceived ease of use and usefulness (Kim and Malhotra 2005), information quality and systems quality (e.g., DeLone and McLean 1992), restrictiveness of the system guidance (Silver 1990), and the trade-off between cognitive effort and guidance (Todd and Benbasat 1999). DeSanctis (1983) uses expectancy theory to suggest that users are more motivated to use DSS if they believe that greater usage will lead to better performance. Our research suggests that users will resist use unless DSS are *designed* to help users understand the basis for the DSS recommendations and how following those recommendations will lead to better performance. We contribute to the information systems/DSS design literature by proposing how "designed feedback" enables users to internalize the rationale underlying DSS recommendations, leading to better evaluations of high-quality model-based DSS. Our work builds on research by Chenoweth et al. (2004) who showed that feedback can help decision makers' transition from less complex DSS models to more complex DSS models. Note that our proposed feedback mechanisms can be implemented by computerized systems which allow for real-time calibration of users' mental models, close to impossible to achieve by any manual method (Sengupta and Te'eni 1993).

We empirically demonstrated that deep learning (i.e., the transformation of mental models) is crucial for managers to form a favorable evaluation of an objectively high-quality DSS. Our study shows that a DSS that provides upside potential feedback can motivate managers to perform better, resulting in greater effort. However, increased effort alone is not sufficient to generate deep learning; the DSS must also

provide clear guidance about how and why a modification of a mental model leads to a superior outcome. Our results also show that mere shallow learning does not lead to better evaluations of the DSS, implying that DSS that offer no opportunity to understand their recommendations are likely to be poorly evaluated by users and, hence, used less frequently. We found, hence, that a *dual-feedback approach*, combining upside potential and specific guidance is required to help managers internalize and be able to take advantage of the DSS model.

We also find that DSS feedback influences users' underlying learning process, which in turn, helps users internalize the relationship between decisions and outcomes. Although much prior research has examined decision outcomes, our study enriches the story by showing how objectively superior DSS can also be perceived more positively. To improve user recognition of DSS value, the DSS should stimulate its users' learning processes by providing "dual-feedback" in an interactive manner. Such feedback is also likely to influence effort-accuracy tradeoff (Todd and Benbasat 1999) in favor of more accurate decisions and better performance. Effective feedback is thus an important driver of DSS evaluations along with variables such as perceived usefulness and ease of use (Davis 1989), incentives (Todd and Benbasat 1999), top management support (Rigby 2001), and training and support (Leonard-Barton and Deschamps 1988).

In summary, our main contributions are: (1) specification of the role of deep learning (i.e., mental model changes) on evaluations of DSS (Figure 3); (2) assessment of the individual and *joint* effects of two types of feedback, corrective and upside, on deep learning in the DSS context, and (3) development of the "3-Gap" framework to understand DSS evaluation (Figure 1). We have also conceptualized and used an unobtrusive mechanism to assess DSS users' mental models and their changes, a methodological approach we hope other researchers will find useful.

### 4.2. Limitations and Future Research

Our work suggests a number of avenues for fruitful future research, drawing from two broad categories of limitations in our study—first, the nature of participants and the context within which they made decisions and, second, the nature of the task and our measurements.

In situations where decision makers have more experience than some of the participants in our study, the former are likely to have strong a priori beliefs that may be hard to overcome through DSS-based learning. An important empirical question is whether DSS can still lead to deep learning when managers either have good intuition (i.e., a small Gap 3) or access to information that cannot easily be incorporated into a decision model (i.e., a small Gap 3 relative to Gap 2). Research of this nature is crucial for the development of an understanding of the role of DSS in bringing about enduring changes to how decisions are made in organizations. We focused especially on the evaluation process and did not include a direct measure of sustained usage, a measure best obtained in a longitudinal field study. Such a longitudinal study would help us understand how feedback affects the dynamics of mental model changes—Does type of feedback affect a users' experimentation over time? How does that experimentation affect the durability of mental model change? We have been neutral about organizational context in our research and an interesting research question is whether and how a DSS can overcome organizational constraints to learning, such as past norms or policies, budget constraints, and hierarchical management structures. Our study focused on understanding the effect of deep learning on DSS evaluations and we did not find any effect of shallow learning. Further study is needed to generate sufficient variation in shallow learning to see how such variation affects DSS evaluations and acceptance. Further studies in different research contexts and with different DSS and preferably also in "real" organizations with "real" DSS should enhance the generalizability of our findings.

In terms of the nature of the task in our study, we provided a single numerical form of feedback (a best guess of what would have happened with the DSS); it would be useful to study alternative feedback mechanisms, including prediction intervals. We assumed our DSS was of very high quality and we did not attempt to vary that quality. An interesting issue is whether the quality of the DSS (measured in terms of Gap 2) affects the efficacy of the feedback mechanisms to reduce Gap 1. While expectancy theory

(Vroom 1964, DeSanctis 1983) suggests that such an effect should occur, further research on the topic is clearly merited. In addition, we assumed the functional form (linear additive) of our participants' mental models to be the same as that of the DSS model. Mental model functional forms might be non-linear and might vary across managers, also leading to interesting research questions. Although participants were asked to rate donors on their attractiveness for solicitation; a more realistic option might have been to ask them to choose donors to solicit, providing binary data for mental model calibration. We also did not provide information to participants as to whether the 20 donors actually made donations; in more realistic settings, such information will likely be available, which complicates the mental model calibration task.

### 4.3. DSS Design and Managerial Implications

For firms that develop and market model-based DSS (e.g., Salesforce.com, DemandTec), our results reinforce the importance of incorporating the two types of feedback in tandem to boost the value that users find in a DSS, thereby increasing the likelihood of use. We note that managers who have internalized the DSS model might not necessarily be able to articulate the weights or functional form of the DSS model; deep learning need not be conscious to be effective.

Our research suggests a way to obtain a better return on investment for firms that have been heavily investing in DSS such as Customer Relationship Management (CRM) and management dashboards, investments whose success has been challenged. For example, Kale (2004) suggests that 60%–80% of CRM investments produce returns well below expectations. Industry analysts suggest that line managers do not necessarily recognize the benefits of systems, leading to increased resistance to adopt, which eventually proves costly for the firm (Petouhoff 2006, p. 6). Roach (2002, as quoted in the IDC report, p. 6) suggests that sustained enhancement in performance results less from technological breakthroughs and more from substantial changes in the "*cerebral production function*" of the knowledge worker. Our results provide direct evidence that investments in DSS that are designed to help transform the mental models (i.e., cerebral production functions) of knowledge workers are more likely to have substantial payoffs. Similarly, in other areas such as medicine and accounting, DSS that incorporate feedback mechanisms in their design are more likely to be used, and thus positively affect user performance.

Overall, we believe we have taken a step toward better understanding of the mental model barriers to DSS use and how DSS can be better designed to overcome those barriers.

## Acknowledgments

The authors thank Professor Ritu Agarwal, Professor Giri Kumar Tayi, and two anonymous reviewers for helpful comments that significantly improved this paper; Sonke Albers, Meg Meloy, Bill Ross, Chuck Weinberg, and Berend Wierenga for constructive feedback on the first version of this paper; and Bob Wood for extensive discussions with the first author on the psychological foundations of this research. The authors also thank the Institute for the Study of Business Markets and the Laboratory for Economics Management and Auctions (both at Pennsylvania State University) for providing financial support and experimental laboratory facilities, respectively.

## References

Agresti, A. 2002. *Categorical Data Analysis*, 2nd ed. John Wiley & Sons, Inc., New York.

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*, 2nd ed. John Wiley & Sons, Inc., New York.

Ashton, R. 1991. Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *J. Accounting Res.* **28** 148–186.

Atkins, P. W. B., R. E. Wood, P. J. Rutgers. 2002. The effects of feedback format on dynamic decision making. *Organ. Behav. Human Decision Processes* **88**(2) 587–604.

Balzer, W. K., L. M. Sulsky, L. B. Hammer, K. E. Sumner. 1992. Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organ. Behav. Human Decision Processes* **53**(1) 35–54.

Bandura, A. 1997. *Self-Efficacy*: *The Exercise of Control*. W. H. Freeman, New York.

Banker, R., R. J. Kauffman. 2004. The evolution of research on information systems: A fiftieth-year survey of the literature in *Management Science*. *Management Sci.* **50**(3) 289–298.

Benbasat, I., A. S. Dexter. 1985. An empirical evaluation of graphical and color-enhanced information presentation. *Management Sci.* **31**(11) 1348–1364.

Chenoweth, T., K. L. Dowling, R. D. St. Louis. 2004. Convincing DSS users that complex models are worth the effort. *Decision Support Systems* **37**(1) 71–82.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* **13**(3) 319–340.

D. Shephard Associates. 1999. *The New Direct Marketing*, 3rd ed. McGraw-Hill, New York.

DeLone, W. H., E. R. McLean. 1992. Information systems success: The quest for the dependent variable. *Inform. Systems Res.* **3**(1) 60–95.

DeSanctis, G. 1983. Expectancy theory as an explanation of voluntary use of a decision support system. *Psych. Rep.* **52** 247–260.

Earley, C. P., G. B. Northcraft, C. Lee, T. R. Lituchy. 1990. Impact of process and outcome feedback on the relation of goal setting to task performance. *Acad. Management J.* **33**(1) 87–105.

Einhorn, H. J., R. M. Hogarth. 1985. Ambiguity and uncertainty in probabilistic inference. *Psych. Rev.* **92**(4) 433–461.

Gentner, D., A. L. Stevens, eds. 1983. *Mental Models*. Lawrence Erlbaum Associates, Mahwah, NJ.

Gonul, M. S., D. Onkal, M. Lawrence. 2006. The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems* **42**(3) 1481–1493.

Goodman, J. S. 1998. The interactive effects of task and external feedback on practice performance and learning. *Organ. Behav. Human Decision Processes* **76**(3) 223–252.

Goodman, J. S., R. E. Wood, M. Hendrickx. 2004. Feedback specificity, exploration, and learning. *J. Appl. Psych.* **89**(2) 248–262.

Hoch, S. J., D. A. Schkade. 1996. A psychological approach to decision support systems. *Management Sci.* **42**(1) 51–64.

Hughes, C. T., M. L. Gibson. 1991. Students as surrogates for managers in a decision-making environment: An experimental study. *J. Management Inform. Sys.* **8**(2) 153–166.

Hunt, D. L., R. B. Haynes, S. E. Hanna, K. Smith. 1998. Effects of computer-based clinical decision support systems on physician performance and patient outcomes. *J Amer. Med. Assoc.* **280**(15) 1339–1346.

IDC. 2002. The financial impact of business analytics. Report, Interactive Data Corporation, Framingham, MA. www.idc.com.

Kale, S. H. 2004. CRM failure and the seven deadly sins. *Marketing Management* **13**(5) 42–46.

Keeney, R. L., H. Raiffa. 1976. *Decision Making with Multiple Objectives*: *Preference and Value Tradeoffs*. John Wiley and Sons, Inc., New York.

Kim, S. S., N. K. Malhotra. 2005. A longitudinal model of continued IS use: An integrative view of four mechanisms underlying post-adoption phenomena. *Management Sci.* **51**(5) 741–755.

Kluger, A. N., A. Denisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. *Psych. Bull.* **119** 254–284.

Kunreuther, H. 1969. Extensions of Bowman's theory on managerial decision-making. *Management Sci.* **15**(8) 415–439.

Lai, F., J. Macmillan, D. H. Daudelin, D. M. Kent. 2006. The potential of training to increase acceptance and use of computerized decision support systems for medical diagnosis. *Human Factors* **48**(1) 95–108.

Leonard-Barton, D., I. Deschamps. 1988. Managerial influence in the implementation of new technology. *Management Sci.* **34**(10) 1252–1265.

Lilien, G. L., A. Rangaswamy, G. H. Van Bruggen, K. Starke. 2004. DSS effectiveness in marketing resource allocation decisions: Reality vs. perception. *Inform. Systems Res.* **15**(3) 216–235.

Lim, K. H., I. Benbasat. 2000. The effect of multimedia on perceived equivocality and perceived usefulness of information systems. *MIS Quarterly* **24**(3) 449–471.

Lim, K. H., L. M. Ward, I. Benbasat. 1997. An empirical study of computer system learning: Comparison of co-discovery and self-discovery methods. *Inform. Systems Res.* **8**(3) 254–268.

Limayem, M., G. DeSanctis. 2000. Providing decisional guidance for multicriteria decision making in groups. *Inform. Systems Res.* **11**(4) 386–401.

Lister, G. J. 2001. *Building Your Direct Mail Program*. Jossey-Bass, San Francisco.

Locke, E. A., K. N. Shaw, L. M. Saari, G. P. Latham. 1981. Goal setting and task performance: 1969–1980. *Psych. Bull.* **90**(1) 125–152.

Mawhinney, C. H., A. L. Lederer. 1990. A study of personal computer utilization by managers. *Inform. Management* **18**(5) 243–253.

McIntyre, S. H. 1982. An experimental study of the impact of judgment-based marketing models. *Management Sci.* **28**(1) 17–33.

Montgomery, A. 2005. The implementation challenge of pricing decision support systems for retail managers. *Appl. Stochastic Models Bus. Indust.* **27**(4–5) 367–378.

Nicolaou, A. L., D. H. McKnight. 2006. Perceived information quality in data exchanges: Effects on risk, trust, and intention to use. *Inform. Systems Res.* **17**(4) 332–351.

Norman, D. A. 1983. Some observations on mental models. D. Gentner, A. L. Stevens, eds. *Mental Models*. Lawrence Erlbaum Associates, Mahwah, NJ.

Parasuraman, A., V. A. Zeithaml, L. L. Berry. 1985. A conceptual model of service quality and its implications for future research. *J. Marketing* **49**(4) 41–50.

Petouhoff, N. 2006. The scientific basis for CRM failures. *Customer Relationship Management Magazine* **10**(3) 48.

Ramaprasad, A. 1987. Cognitive process as a basis for MIS and DSS design. *Management Sci.* **33**(2) 139–148.

Reda, S. 2003. Despite early positive results, retailers haven't jumped on analytics bandwagon. *Stores* **85**(3) 34.

Remus, W. 1996. Will behavioral research on managerial decision making generalize to managers? *Managerial Dec. Econom.* **17**(1) 93–101.

Rigby, D. 2001. Management tools and techniques: A survey. *California Management. Rev.* **43**(2) 139–160.

Sanders, N., K. B. Manrodt. 2003. Forecasting software in practice: Use, satisfaction, and performance. *Interfaces* **33**(5) 90–93.

Schlegelmilch, B., A. Diamantopoulos, A. Love. 1997. Characteristics affecting charitable donations: Empirical evidence from Britain. *J. Marketing Practice*: *Appl. Marketing Sci.* **3**(1) 14–28.

Sengupta, K., D. Te'eni. 1993. Cognitive feedback in GDSS: Improving control and convergence. *MIS Quart.* **17**(1) 87–113.

Shim, J. P., M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, C. Carlsson. 2002. Past, present, and future of decision support technology. *Decision Support Systems* **33**(2) 111–126.

Sieck, W. R., H. R. Arkes. 2005. The recalcitrance of overconfidence and its contribution to decision aid neglect. *J. Behavioral Decision Making* **18**(1) 29–53.

Silver, M. S. 1990. Decision support systems: Directed and nondirected change. *Inform. Systems Res.* **1**(1) 47–70.

Silver, M. S. 1991. Decision guidance for computer based decision support. *MIS Quart.* **15**(1) 105–122.

Singh, D. T., P. P. Singh. 1997. Aiding DSS users in the use of complex OR models. *Ann. Oper. Res.* **72** 5–27.

Sintchenko, V., E. Coiera, J. Iredeli, G. Gilbert. 2004. Comparative impact of guidelines, clinical data, and decision support on prescribing decisions: An interactive Web experiment with simulated cases. *J. Amer. Med. Inform. Assoc.* **11**(1) 71–77.

Sullivan, L. 2005. Fine-tuned pricing. *Inform. Week* (August). www.informationweek.com.

Todd, P., I. Benbasat. 1999. Evaluating the impact of DSS, cognitive effort, and incentives on strategy selection. *Inform. Systems Res.* **10**(4) 356–374.

Tversky, A., D. Kahneman. 1987. Rational choice and the framing of decisions. R. M. Hogarth, M. W. Reder, eds. *Rational Choice*. The University of Chicago Press, Chicago. 251–278.

Udo, J. G., S. J. Davis. 1992. A comparative analysis of DSS user-friendliness and performance. *Internat. J. Inform. Management* **12**(3) 209–223.

Umanath, N. S., I. Vessey. 1995. Multiattribute data presentation and human judgment: A cognitive fit perspective. *Dec. Sci.* **25**(5) 795–824.

Van Bruggen, G. H., A. Smidts, B. Wierenga. 1996. The impact of the quality of a marketing decision support system: An experimental study. *Internat. J. Res. Marketing* **13**(4) 331–343.

Vessey, I., D. Galletta. 1991. Cognitive fit: An empirical study of information acquisition. *Inform. Systems Res.* **2**(1) 63–84.

Vroom, V. H. 1964. *Work and Motivation*. John Wiley & Sons, Inc., New York.

Wilson, J. R., A. Rutherford. 1989. Mental models: Theory and application in human factors. *Human Factors* **31**(6) 617–634.

Wood, R. E., A. Bandura, T. Bailey. 1990. Mechanisms governing organizational performance in complex decision-making environments. *Organ. Behav. Human Decision Processes* **46**(2) 181–201.