NORTHERN ARIZONA UNIVERSITY
The W. A. Franke College of Business

Repeated Task Performance and Numeric Outcome Feedback in a Complex Decision Environment

Working Paper Series—10-09 | May 2010

Steve Buchheit, Texas Tech University*

Derek Dalton, Texas Tech University

Tom Downen, Northern Arizona University

Sonja Pippin, University of Nevada Reno

*Corresponding author:
(806) 742-1542; MS2101, Rawls College of Business, Lubbock TX 79409

Repeated Task Performance and Numeric Outcome Feedback in a Complex Decision Environment

I. INTRODUCTION

Many business decision scenarios are repetitive and involve feedback between decision judgments (Sprinkle, 2000). One particular type of feedback that is frequently offered in business contexts is numeric outcome feedback (OFB). OFB provides individuals with knowledge of quantified results with intuitive directional properties (e.g., 'more profit' is intuitively good and 'more cost' is intuitively bad in most business settings). Relative to other forms of feedback such as explanatory feedback (i.e., step-by-step feedback regarding why a particular answer was 'correct'), OFB is less time-consuming and less costly to deliver (Bonner and Walker, 1994). In a world characterized by time constraints and "information overload" (e.g., Epstein, 2007), exploring decision makers' ability to 'self learn' based on numeric outcome feedback is both (1) increasingly important and (2) naturally related to accounting information.

In this study, we are motivated by the preceding tension between research streams. Specifically, prior psychology research indicates that OFB is largely ineffective at improving performance in complex tasks (see Balzer et al., 1989 for a review). Some psychology studies even suggest that OFB may hinder the learning process in complex tasks, thus leading to declines in judgment performance (Todd and Hammond, 1965; Hammond and Summers, 1972; Hammond et al., 1973). Other research, however, speculates that these prior psychology findings may result from the fact that psychology experiments often involve abstract tasks in which the decision cues (i.e., pieces of information) and criterion (i.e., the predictor variable) have no real-life referents (Libby, 1981; Hirst et al., 1999). In more intuitive decision environments in which the decision cues and criterion have real-life properties (e.g., accounting information), judges may be better able to learn from OFB in complex tasks. Consequently, it is uncertain whether OFB can improve performance in complex tasks which are placed in typical business environments. To investigate this tension, we examine the effectiveness of numeric OFB using a setting developed by Luft and Shields (2001) – a setting characterized as a complex accounting task environment (Elliot et al. 2007).

In contrast to prior psychology findings, our results suggest that numeric OFB can, indeed, improve judgment performance relative to mere repetition (i.e., no feedback) in a complex task. This finding is important given the widespread use of numeric OFB within common accounting tasks, many of which are arguably complex. Consistent with assertions by Libby (1981) and Hirst et al. (1999), when tasks are placed in more meaningful decision environments, individuals appear better able to learn from OFB because the cues and criterion have real-life referents, thus facilitating the learning process. In short, we find that meaningful feedback aids learning and performance while repetition (alone) does not.

Importantly, the Luft and Shields decision environment offers both diagnostic and non-diagnostic cues to decision makers. In complex environments characterized by 'excess' information, finding that summary OFB can substantially improve judgment performance after only one round of feedback is evidence that decision makers quickly learn to 'sort' the diagnosticity of information cues.

In contrast, most prior research investigating the effectiveness of OFB has focused on decision cues which were always related, to some extent, to the measured performance criterion. Many learning environments, however, include cues that are entirely unrelated to the criterion, thus making natural decision judgments more difficult. Our findings indicate that OFB can improve performance in these types of learning environments (i.e., decision-makers effectively 'sort' information diagnosticity).

Our findings lend support to prior OFB-based accounting studies which also find that OFB improves performance (Leung and Trotman, 2005; Hirst et al., 1999; Bonner and Walker, 1994). Importantly, we extend this line of research to a complex setting involving a combination of diagnostic and non-diagnostic decision cues and capture potentially surprising evidence regarding the speed with which decision makers can effectively 'sort' cues. We argue that future research involving summary level OFB and the limits of 'cue sorting' is increasingly important in 'flattened' organizational environments where fewer managers govern larger numbers of subordinates (thus making less time available for direct mentoring or step-by-step feedback).

The remainder of the paper is structured as follows: the next section summarizes the relevant prior literature and derives hypotheses, section three discusses the methodology, section four presents the results, and section five concludes.

II. PRIOR LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

In natural business environments, decisions can involve single iteration judgment tasks such as merger and acquisition transactions, the incorporation of a business, and the change from LIFO to FIFO accounting. Likewise, business decisions can also involve repeated judgment tasks such as buy/sell decisions, the budgeting of labor or material resources during regular operations, and the application of audit procedures when making audit risk assessments. Sprinkle (2000) notes that many managerial decisions are repeated as managers attempt to maximize revenues. When business decisions are repetitious, there is an opportunity for feedback to be provided or obtained between iterations. According to Balzer, et al. (1989, pg 412), "feedback is the process by which an environment returns to individuals a portion of the information in their response output necessary to compare their present strategy with a representation of an ideal strategy…."

While prior psychology research has identified several different kinds of feedback[1] (see Balzer et al., 1994; Kluger and DeNisim 1996; Todd and Hammond 1965; Hammond, et al. 1973), this study focuses on outcome feedback (OFB). OFB provides individuals with the correct answers to particular judgment trials. For example, after completing tax returns, tax professionals typically receive tax reviewer notes as a source of outcome feedback. Other examples of OFB within the accounting domain include prior years' audit/tax work papers, audit reviewer notes, and financial statements that compare actual data to budgeted data.

Bonner and Walker (1994) find that explanatory feedback[2] (i.e., step-by-step feedback regarding why a particular answer was correct) improves auditor performance in a ratio-analysis task more than OFB (i.e., the correct answer). Psychology research also supports this finding (see Balzer et al., 1989). With explanatory feedback, individuals are informed why a particular outcome occurred; with OFB, however, individuals have to work backwards from the outcome to infer why a particular outcome occurred. Consequently, explanatory feedback allows for a more simplified learning process, which, in turn, leads to greater improvements in judgment performance.

Despite the advantages of explanatory feedback, Bonner and Walker (1994) note that providing explanatory feedback may be unfeasible in reality. For example, oftentimes audit seniors are too busy to provide explanatory feedback to novice auditors (Earley, 2001). OFB, in contrast, is less costly and time-consuming (Bonner and Walker, 1994). Further, depending on the circumstances, OFB may be the only type of feedback that is offered, especially when conditions involve time constraints (Bonner and Walker, 1994). Hirst et al. (1999) note that OFB is an important part of any review process (e.g., tax reviews, audit reviews, etc.) whereby a reviewer informs a preparer of any incorrect preparer-based judgments. Given the importance of OFB as a low-cost form of feedback within a wide range of accounting domains, this

[1] In addition to OFB, prior psychology research also identifies cognitive feedback (CFB) and task properties feedback (TPF). CFB provides individuals with information regarding their own judgment policies (Balzer et al., 1989; Leung and Trotman 2005). Judgment policies refer to the manner in which individuals use information cues (i.e., various pieces of information) to arrive at judgments. Individuals may be unaware of their judgment policies, especially in the case of complex tasks; thus, CFB informs decision makers of these processes. Task-properties feedback provides individuals with information related to the optimal weights for decision cues in a task. Essentially, task-properties feedback instructs individuals on how to optimally use decision cues to improve judgments.

[2] Explanatory feedback includes both outcome feedback (i.e., the correct answer) and task properties feedback (i.e., an explanation for why the outcome occurred).

paper investigates the effectiveness of OFB. In particular, this study examines the effectiveness of OFB in a relatively complex accounting task.

Prior psychology research indicates that OFB is ineffective at improving performance in complex tasks (see Balzer et al., 1989 for a review). Further, Kluger and DeNisi (1996) perform a meta-analysis of research examining OFB and find that for 38% of the studies in their sample, OFB actually lead to declines in performance. Kluger and DeNisi suggest that while OFB may be beneficial with simple tasks, it is largely ineffective in complex tasks. Some psychologists even posit that OFB may actually impede the learning process in complex tasks, thus leading to lower performance (Todd and Hammond 1965; Hammond and Summers, 1972; Hammond et al. 1973).

Hirst et al., (1999) and Libby (1981) suggest that the ineffectiveness of OFB in prior psychology studies may exist because psychology research tends to use simulated, abstract tasks. In more meaningful decision environments in which participants have knowledge about the relative importance of the decision cues and the relationships between decision cues and the criterion (the predictor of the cues), OFB may lead to improved judgments (Libby, 1981; Hirst et al., 1999). Consistent with this assertion, prior accounting research which uses non-abstract, realistic tasks has found that OFB can, indeed, improve performance[3] (Hirst and Luckett, 1992; Hirst et al., 1999; Leung and Trotman, 2005; Earley, 2003; Nelson, 1993; Ashton, 1990; Harrell, 1977).

Bonner and Walker (1994) suggest that for OFB to be effective, the task must be simple enough for participants to work backwards from the outcome (i.e., the criterion) to gain an understanding for why the outcome occurred (i.e., the relationship between the cues and the criterion). Simple tasks have high predictability, relatively few decision cues, or are simplified in other ways (Bonner and Walker, 1994). For example, several prior OFB-based accounting studies use tasks in which the criterion can be predicted with perfect accuracy (Hirst and Luckett, 1992; Harrell, 1977). Other accounting studies which study OFB use simplified tasks with only a few decision cues (Nelson, 1993; Ashton, 1990; Leung and Trotman, 2005). Consequently, most of the prior OFB-based accounting studies have been classified as relatively simplistic (Bonner and Walker, 1994). As such, the effect of OFB in complex accounting tasks remains uncertain.

For OFB to be effective, participants must be able to work backwards from the outcome (i.e., the criterion) to gain an understanding for why the outcome occurred (i.e., the relationship between the cues and the criterion). However, this does not necessarily suggest that tasks must be simplistic for OFB to be effective. For example, Hirst et al. (1999) manipulate task complexity and find that even in tasks that are not perfectly predictable, OFB is still effective at improving performance. Yet, Hirst et al. still classify their task as relatively simple and encourage future research to examine the effectiveness of OFB in more complex accounting-based tasks.

To test whether OFB can improve performance in a complex task, we investigate the effectiveness of OFB in an accounting scenario which has been classified as high in complexity[4] (Elliott, et al. 2007). Elliott et al. (2007) provide a summary of recent behavioral research publications in prominent accounting journals and categorize each of the experimental tasks employed therein into whether each was low or high in integrative complexity. Table 1 lists the studies that they reviewed and includes a notation about whether the study involved a single task iteration or repeated tasks.

---

[3] Hirst and Luckett (1992) find that OFB improves staff evaluations; Hirst et al. (1999) find that OFB improves bankruptcy predictions; Leung and Trotman (2005) find that OFB improves misstatement risk assessments in a non-configural (i.e., linear) task; Earley (2003) finds that OFB increases the level of reasoning, which, in turn, improves discount rate evaluations; Nelson (1993) finds that OFB reduces financial statement errors; Ashton (1990) finds that OFB improves bond rating evaluations; and Harrell (1977) finds that OFB improves performance evaluations.

[4] Elliott, et al. (2007) classify experimental tasks as high in integrative complexity if participants must draw "relatively complex connections among distinct pieces of information provided for the purpose of making investment-related judgments and decisions" (p 36).

Table 1. Summary of Studies Reviewed by Elliott, et al. (2007)

Publication	Integrative Complexity Level	Single or Repeated Task
Hirst, Koonce, and Simko (1995)	Low	Single
Bloomfield and Libby (1996)	Low	Repeated
Maines and Hand (1996)	Low	Repeated
Kennedy, Mitchell, and Sefcik (1998)	High	Single
Lipe (1998)	Low	Single
Bloomfield, Libby, and Nelson (1999)	Low	Repeated
Hirst, Koonce, and Miller (1999)	Low	Repeated
Bloomfield and Wilks (2000)	Low	Repeated
Maines and McDaniel (2000)	High	Single
Hodge (2001)	Low	Single
-- Luft and Shields (2001) ---------------	**High**	**Single**
Bloomfield, Libby, and Nelson (2003)	Low	Repeated
Hirst, Jackson, and Koonce (2003)	High	Single
Nelson, Krische, and Bloomfield (2003)	Low	Repeated
Frederickson and Miller (2004)	Low	Single
Hodge, Kennedy, and Maines (2004)	High	Single
Krische (2005)	High	Repeated
Mercer (2005)	Low	Single
Elliott (2006)	Low	Single

To assess whether OFB can improve performance in a complex task, we use the task employed by Luft and Shields (2001; hereafter L&S). In brief, the task provides participants with intangible expenditures for twenty similar manufacturing plants. Participants must use four decision cues (intangible expenditures for the current and three previous periods) to predict the current period profits. There is no statistically significant relationship between the current period profits (i.e., the criterion) and the current or two previous periods' expenditures. Only the lagged intangible expenditures (i.e., intangible expenditures from three periods prior to the current period) are related to current profits. Consequently, only one of the four cues is related to the criterion, thus adding additional noise and complexity to the task.

As an additional contribution, the relatively complex L&S task allows us to examine additional boundary conditions under which OFB may improve performance. Hirst et al. (1999), extends the conditions under which OFB improves performance (by showing that OFB improves performance in a task that is not perfectly predictable[5]). Similarly, the L&S task allows us to extend prior accounting

[5] The task predictability in one of the tasks used by Hirst et al. (1991) equals .81. The task predictability used in our study equals .88. Hirst et al. (1999) define task predictability as "the extent to which the cues relate to and are used to predict the environmental event" (p. 289). In other words, if a participant were to perfectly apply the information

studies which examine OFB. In prior OFB-based accounting research, decision cues were always related, to some extent, to the criterion under investigation[6]. In the L&S task, however, only the most distant cue—the lagged decision cue—is related to the criterion, thus adding an extra element of complexity not previously considered.

Tasks that associate each cue with the criterion may oversimplify many real-life learning environments. Conceivably, many learning environments include cues that are entirely unrelated to the criterion. Consequently, the L&S task will address whether OFB can still improve judgments in these types of learning environments.

Despite prior psychology research that suggests that OFB is either ineffective or even counter-effective in complex experimental tasks (Balzer et al., 1989; Todd and Hammond 1965; Hammond et al. 1973), we rely upon prior accounting findings which find that OFB is effective in more meaningful decision environments and predict that OFB will improve performance even in a relatively complex accounting task.

H1: Providing decision makers with outcome feedback will improve decision performance in a complex task.

Learning models assume that learning only occurs after participants receive feedback on prior performance (see Erev and Roth, 1998 and Cheung and Friedman, 2003). Several experimental economics studies, however, find that learning can occur even in the absence of feedback (Weber, 2003; Rapoport et al., 2002; Grether, 1980). These studies suggest that learning can exist in no-feedback environments because of individual introspection. For example, the contemplation that occurs from making multiple judgments may lead to an improved understanding of the underlying environmental relationships or perhaps a fortunate discovery of relevant task principles. Without taking into account the learning that may occur from mere repetition alone, the overall importance of OFB may be overstated. To account for this possibility, we include a condition in which participants complete two iterations of the experimental task without receiving OFB between iterations.

In contrast to the aforementioned experimental economics research, prior accounting research finds that mere repetition alone does not lead to improvements in judgment performance (Leung and Trotman, 2005; Hirst et al., 1999; Bonner and Walker, 1994). Despite these inconsistencies, both economics and accounting research consistently finds that OFB improves performance incrementally more than mere repetition alone (Hirst and Luckett, 1992; Bonner and Walker, 1994; Earley, 2001; Hirst et al., 1999; Leung and Trotman, 2005; Weber, 2003), leading to the following hypothesis:

H2: Providing decision makers with outcome feedback will improve decision performance in a complex task incrementally more than simply allowing repetition.

As an additional contribution, this study attempts to replicate L&S's fixation effect. Bamber et al. (2000) argue that while replications in accounting are infrequent, they are critical to gaining a fuller understanding of the conditions under which research findings either persist or diminish. They note that accounting researchers typically do not fully appreciate that prior findings are likely sensitive to research design choices, thus leading to overgeneralizations of seminal findings. For example, they find that Beaver's seminal 1968 article is much more sensitive to research design choices than typically

---

provided in the environmental model, a perfectly predictable task would allow the participant to make perfect judgments (see Hirst et al. (1999) for further detail).

[6] For example, in Hirst et al. (1999) three financial ratios were each related to the likelihood of firm bankruptcy; in Hirst and Luckett (1999), five job-related cues were each related to staff evaluations; in Leung and Trotman (2005), five audit procedures were each related to misstatement risk assessments; in Nelson (1993), three financial cues were each related to financial statement errors; and in Ashton (1990), three financial cues were each related to bond ratings.

interpreted[7]. Without replications, however, the fragility or robustness of prior important findings remains uncertain and overgeneralizations are likely to persist. We examine whether a different set of participants (i.e., MSA students rather than MBA students) influences the robustness of L&S's findings.

L&S' premise is that fixation on accounting terms ("investment" vs. "expense") influences an individual's ability to see the relationship between an expenditure made three quarters ago and gross profits in the current quarter. More specifically, L&S find that when an intangible expenditure is classified as an investment (expense), judgments are more (less) accurate. In essence, fixation on the word investment (expense) improves (worsens) an individual's ability to see the lagged expense-profit relationship. This finding is important because it suggests that accounting can influence individuals' focus of attention, which, in turn, affects the learning process.

While L&S use Master of Business Administration (MBA) students as participants, we use Master of Science in Accounting (MSA) students. Given that MSA students take more accounting classes than MBA students, MSA students may have a more sophisticated understanding of accounting terminology. Consequently, MSA students may be better able to see through the labels, thus potentially mitigating or even eliminating the fixation effect found in L&S. Initially, however, we expect to replicate the fixation effect found in L&S:

> *H3: Decision performance will be better when intangible expenditures are reported as investments rather than expenses (fixation effect).*

III. METHODOLOGY

Experiment. First, we attempt to replicate the original L&S study, which includes exposure to learning data and then a single iteration of participant task performance.[8] Then we extend the L&S study by adding a second iteration of participant performance. Consistent with L&S, our participants were provided with performance-based compensation in exchange for their participation in the experiment.

Replication. Ninety volunteer Master of Science in Accounting (MSA) students completed a pen-and-paper experimental task. They received information about twenty manufacturing plants (all within a single company) that produced the same product and used the same technology. Participants were informed that the company had recently implemented a quality improvement program and that the company was trying to determine what effect, if any, discretionary quality improvement spending would have on the gross profits of each plant. To the extent the quality improvement spending did affect gross profit, participants were told that the effect should be comparable across plants. Participants were provided with "learning data" for each of the twenty plants. This data included actual quality improvement spending for the most recently completed quarter and the three preceding quarters, as well as the actual gross profit for the most recently completed quarter (similar to Table 2, with profit amounts included) for each of the twenty plants. Using this learning data, participants attempted to determine the relationship that existed between the quality improvement expenditures and profitability. As shown in Table 2, the between-participants manipulation in L&S is the labeling of the quality improvement expenditures as either expense or investment. After studying the learning data, participants received quality improvement spending data for twenty similar plants and were asked to predict gross profit for each plant. All of the experimental materials that we used in the learning phase and in the first task iteration are identical to the materials used by L&S.

---

[7] Bamber et al. (2000) find that, contrary to prior interpretations of Beaver (1968), there is no evidence that most earning announcements have information content. In fact, the market reaction to earnings announcements is driven by only a small fraction of earnings announcements. Bamber et al. note that Beaver's decision to only select non-12/31 firms with fewer than 20 WSJ items in a given year led to a smaller than average sample, which resulted in Beaver's significant research findings.

[8] We thank Joan Luft for providing her experimental instrument. We discuss the experimental task using our second task iteration (Table 2); learning data and data for the first task iteration are discussed fully in pages 570–573 of Luft and Shields (2001). For a quick look at data from the original study, see Tables 1 and 2 of that article.

Table 2. Data Reported for the Second Task Iteration

Plant	Actual quality improvement _______,[a] 3 quarters ago	Actual quality improvement _______,[a] 2 quarters ago	Actual quality improvement _______,[a] previous quarter	Actual quality improvement _______,[a] quarter just completed	Predicted gross profit, quarter just completed
101	$ 2.393 M	$ 1.614 M	$ 1.039 M	$ 1.899 M	
102	$ 0.855 M	$ 1.047 M	$ 1.084 M	$ 1.392 M	
103	$ 1.602 M	$ 0.618 M	$ 2.451 M	$ 1.116 M	
104	$ 1.641 M	$ 0.237 M	$ 1.558 M	$ 0.623 M	
105	$ 1.604 M	$ 2.537 M	$ 2.232 M	$ 1.498 M	
106	$ 1.266 M	$ 2.270 M	$ 0.979 M	$ 0.205 M	
107	$ 1.070 M	$ 1.732 M	$ 2.115 M	$ 0.757 M	
108	$ 0.704 M	$ 1.579 M	$ 1.675 M	$ 1.636 M	
109	$ 1.313 M	$ 1.966 M	$ 0.763 M	$ 0.945 M	
110	$ 1.342 M	$ 0.270 M	$ 1.094 M	$ 0.719 M	
111	$ 2.683 M	$ 0.606 M	$ 0.613 M	$ 1.548 M	
112	$ 1.004 M	$ 1.700 M	$ 0.723 M	$ 0.827 M	
113	$ 2.209 M	$ 1.641 M	$ 0.972 M	$ 0.644 M	
114	$ 2.054 M	$ 1.652 M	$ 0.789 M	$ 1.039 M	
115	$ 1.632 M	$ 2.087 M	$ 0.875 M	$ 1.325 M	
116	$ 0.668 M	$ 1.836 M	$ 1.463 M	$ 1.128 M	
117	$ 0.879 M	$ 1.702 M	$ 0.660 M	$ 0.997 M	
118	$ 2.164 M	$ 1.352 M	$ 1.111 M	$ 0.859 M	
119	$ 0.814 M	$ 0.805 M	$ 0.655 M	$ 2.206 M	
120	$ 2.221 M	$ 2.002 M	$ 1.446 M	$ 2.101 M	

[a] *The term "expense" or "investment" was inserted in the blank, depending on the experimental condition of the applicable participant.*

Extension. Fifty-two of the participants were subsequently provided with the actual gross profit results for each plant in the first task iteration, which they could then compare with their predicted gross profit for each plant (representing OFB). To differentiate between the effects of repeated task performance and OFB, the other 38 participants did not receive any OFB after the first task iteration (but did repeat the task).9 In our second task iteration, participants were presented with data for twenty additional plants and

[9] In addition to manipulating whether OFB was provided between the two task iterations, we also included in our study a manipulation of the typeface used for a portion of the data in the second task iteration. For one participant group (n = 26), all of whom received OFB, we bolded the column representing the quality improvement spending three quarters prior to the most recently completed quarter. This manipulation was introduced in an attempt to establish an upper bound on potential learning in the second task iteration. We expected that the bold manipulation would, in effect, create a control group. Due to the probabilistic and predictive nature of the task, no absolute correct

Table 3. Correlations between Reported Data Elements

Panel A. Learning Data

		Quality Spending			
	Gross Profit $_t$	t	t-1	t-2	t-3
Gross Profit $_t$	1.00	-0.27	-0.02	-0.17	0.90*
Quality Spending $_t$		1.00	-0.23	-0.12	-0.18
Quality Spending $_{t-1}$			1.00	0.30	0.01
Quality Spending $_{t-2}$				1.00	-0.12
Quality Spending $_{t-3}$					1.00

Panel B. First Task Iteration

		Quality Spending			
	Gross Profit $_t$	t	t-1	t-2	t-3
Gross Profit $_t$	1.00	-0.36	0.17	0.07	0.95*
Quality Spending $_t$		1.00	-0.33	0.09	-0.28
Quality Spending $_{t-1}$			1.00	-0.03	0.13
Quality Spending $_{t-2}$				1.00	0.04
Quality Spending $_{t-3}$					1.00

Panel C. Second Task Iteration

		Quality Spending			
	Gross Profit $_t$	t	t-1	t-2	t-3
Gross Profit $_t$	1.00	0.05	0.13	-0.02	0.91*
Quality Spending $_t$		1.00	-0.00	0.01	0.11
Quality Spending $_{t-1}$			1.00	0.03	-0.11
Quality Spending $_{t-2}$				1.00	-0.07
Quality Spending $_{t-3}$					1.00

** correlation differs significantly from zero ($p < 0.05$)*

were asked to again provide gross profit predictions (as shown in Table 2). The data for the third group of plants was generated using the same general algorithm that created the learning data and the data for the first task iteration (i.e., we use the same data generating process as was employed by L&S).

Table 3 provides the correlation matrices for each of the three data sets utilized in this study. (Panels A and B replicate L&S. Panel C is new to the current study.) As in the L&S data, the actual gross profit data in the second task iteration is most closely correlated with the quality improvement spending in

---

answers were expected. However, the responses provided in the bolded condition should have represented the best possible improvement, since individuals in this condition were provided with a strong cue highlighting the relationships between the data components. The results showed that the bolding manipulation had little effect, mainly because even those without the bolding improved greatly when provided with OFB.

the earliest quarter presented (three quarters prior to the most recently completed quarter). The three-quarter lagged quality improvement spending and current gross profit have a correlation of 0.90 in the learning data, 0.95 in the first task iteration, and 0.91 in the second task iteration (all three differ significantly from zero, $p < .05$). The correlations between the quality improvement spending amounts in the other quarters and the gross profit in the most recently completed quarter, as well as the correlations among the separate quality improvement spending amounts, are not statistically different from zero.

IV. RESULTS

H1 addresses whether OFB will improve performance, measured as lower average absolute prediction errors[10] (APEs) in the second task iteration, and H2 addresses whether OFB improves performance incrementally more than merely repeating the task. As shown in Table 4, Panel B, the APEs decrease from $4.10 million in the first iteration to $3.56 million in the second iteration ($p = .001$). This decrease in prediction error is significant for both the expense ($p = .03$) and investment ($p = .02$) groups. However, these results do not explain whether task repetition or OFB or both cause this improvement.

To better evaluate H1 and H2, we separate our participants based on whether they received OFB before completing the second task iteration. Approximately half of our participants in both the expense and the investment group received OFB before completing the second task iteration.[11] As shown in Panel C of Table 4, the improvement caused by OFB is significant ($p < .001$), suggesting that OFB is capable of improving performance in a complex task. Further, the group that did not receive OFB did not exhibit significant improvement in average prediction errors over the two iterations ($p = .79$), suggesting that mere repetition does not improve performance in our complex task. These findings support both H1 and H2. Figure 1 displays this relationship graphically.

With regards to H3, Table 4, Panel A compares the average absolute prediction error (APE) results from the L&S study to the results from the replication portion of our study. L&S's participants have average APEs of $4.81 million (expense condition) and $3.94 million (investment condition). Performing the same task, our participants' average APEs are $4.16 million (expense condition) and $4.05 million (investment condition). Thus, similar to the L&S study, our investment participants performed better (lower APEs) than our expense participants; however, in our case, the difference is not significant. As such, we fail to replicate the fixation effect identified in L&S.[12] Specifically, L&S find that fixation on the short-term implications of the expense label causes participants to overweight potential short-term lags between quality improvement expenditures and gross profit while simultaneously underweighting long-term (and, in this case, correct) lags between expenditures and gross profit (significant in the L&S study at $p = .06$; insignificant in our study at $p = .75$; two-tailed tests). Consequently, we fail to support H3.

[10] For each of the twenty plants, predicted profit could be higher or lower than realized profit. As such, absolute prediction error is averaged over the twenty plants.

[11] Twenty-five of the 44 (57%) participants in the expense condition received OFB. Twenty-seven of the 46 (59%) participants in the investment condition received OFB.

[12] One notable difference between the L&S study and the first iteration of our study is the type of participants used. The L&S participants were Master of Business Administration (MBA) students, whereas we used Master of Science in Accounting (MSA) students. We assert that one major cause for our expense participants performing better (lower APEs) than the L&S expense participants, which was the cause for the lack of a significant fixation effect in our study, is that MSA students may have a more sophisticated understanding of accounting terms than do MBA students. Survey data that we collected separately (untabulated), using different groups of MSA students and MBA students at the same university, support this assertion. This finding (or lack thereof) emphasizes the importance of replicating behavioral accounting studies, to further validate robustness.

Table 4. Average Absolute Prediction Errors (APEs)

Panel A. Replication Results

	Expense Group Mean APE (*std dev*)	Investment Group Mean APE (*std dev*)	*t*-statistic	*p*-value (two-tailed)
L&S (*n* = 37)	$ 4.81 M (*$ 1.22 M*)	$ 3.94 M (*$ 1.43 M*)	1.98	0.06
Current Study (*n* = 90)	$ 4.16 M (*$ 1.76 M*)	$ 4.05 M (*$ 1.39 M*)	0.31	0.75

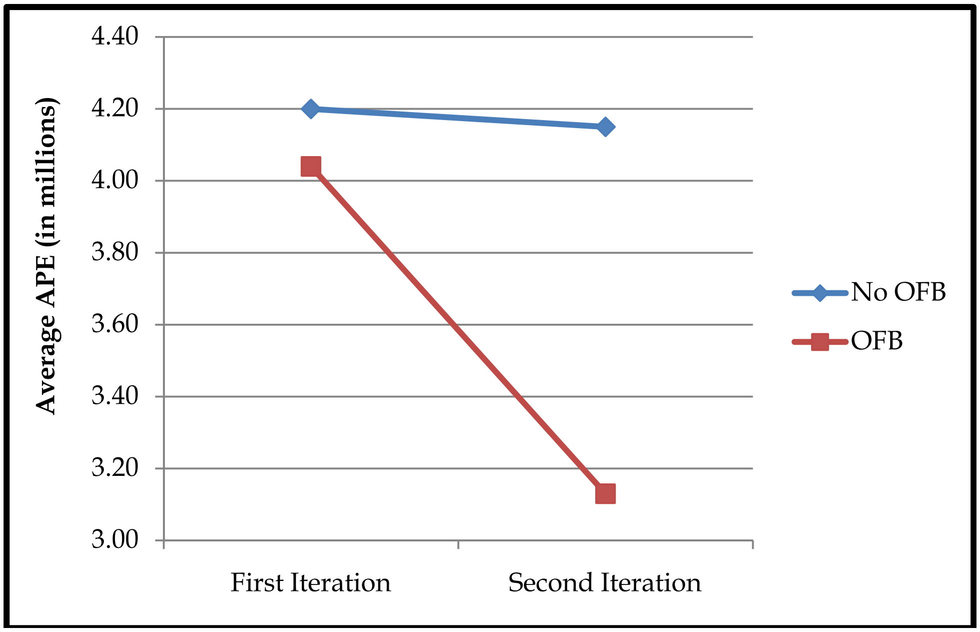
Panel B. First Iteration vs. Second Iteration

	Total Mean APE (*std dev*)	Expense Group Mean APE (*std dev*)	Investment Group Mean APE (*std dev*)	*t*-statistic (*p*-value)
First Iteration	$ 4.10 M (*$ 1.57 M*)	$ 4.16 M (*$ 1.76 M*)	$ 4.05 M (*$ 1.39 M*)	0.31 (0.75)
Second Iteration	$ 3.56 M (*$ 1.21 M*)	$ 3.60 M (*$ 1.17 M*)	$ 3.52 M (*$ 1.27 M*)	0.31 (0.76)
t-statistic (*p*-value)	3.35 (0.001)	2.21 (0.03)	2.53 (0.02)	

Panel C. Outcome Feedback by Iteration

	Total Mean APE (*std dev*)	Outcome Feedback Group Mean APE (*std dev*)	No Outcome Feedback Group Mean APE (*std dev*)	*t*-statistic (*p*-value)
First Iteration	$ 4.10 M (*$ 1.57 M*)	$ 4.04 M (*$ 1.85 M*)	$ 4.20 M (*$ 1.11 M*)	0.47 (0.64)
Second Iteration	$ 3.56 M (*$ 1.21 M*)	$ 3.13 M (*$ 1.02 M*)	$ 4.15 M (*$ 1.23 M*)	4.31 (<0.001)
t-statistic (*p*-value)	3.35 (0.001)	3.74 (<0.001)	0.26 (0.79)	

Figure 1. Interaction Effects between Task Iteration and Outcome Feedback



Recognizing the repeated-task design of our study, we also analyze our results using a repeated-measures regression model to control for potential individual differences that could contribute to performance and learning (and could bias OLS results). This general linear model allows us to test all of the hypotheses in a parsimonious manner, with the results provided in Table 5.

$$APE = \beta_0 + \beta_1 Expense + \beta_2 Iteration + \beta_3 OFB + \beta_4 Iteration \times OFB + \varepsilon$$

where *APE* is the average absolute prediction error for each participant

Expense equals 0 if intangible expenditures are labeled as investments and 1 if intangible expenditures are labeled as expenses

Iteration equals 0 for the first task iteration and 1 for the second task iteration

OFB equals 0 if the participant is not presented with OFB before the second task iteration and 1 if the participant is presented with OFB before the second task iteration

*The regression results are generally consistent with the *t*-tests.

The variable *Iteration* is significant and negative ($p = .004$), indicating an overall reduction of average APEs when repeating the task. Also, the variable *OFB* is significant, suggesting that provision of OFB reduces prediction errors across both iterations.[13] This result provides support for H1. The variable *Expense* is not significant, indicating the lack of fixation ($p = .74$) across both iterations. Consequently, we fail to support H3.

The *Iteration* × *OFB* interaction term indirectly tests H2. This interaction term essentially assesses whether participants who receive OFB improve decision performance incrementally more (as measured by a larger reduction in APEs) than those who do not receive OFB. The *Iteration* × *OFB*

[13] Obviously, there is no reason to expect that providing OFB between iterations would have any effect on performance in the first iteration. However, the highly significant improvement that participants who received OFB experienced in the second iteration drives an overall significant difference for this main effect.

Table 5. Regression Results Using a Repeated Measures General Linear Model

$$\text{MODEL: } APE = \beta_0 + \beta_1\, Expense + \beta_2\, Iteration + \beta_3\, OFB + \beta_4\, Iteration \times OFB + \varepsilon_i$$

Variable	*F*-Value	*p*-value (two-tailed)
Intercept (β_0)	996.25	<0.01
Expense (β_1)	0.11	0.74
Iteration (β_2)	8.91	<0.01
OFB (β_3)	5.72	0.02
Iteration × *OFB* (β_4)	7.36	0.01

APE: the average absolute prediction error for each participant

Expense: equals 0 if intangible expenditures are labeled as investments and 1 if intangible expenditures are labeled as expenses

Iteration: equals 0 for the first task iteration and1 for the second task iteration

OFB: equals 0 if the participant is not presented with outcome feedback during the second task iteration and 1 if the participant is presented with outcome feedback during the second task iteration

interaction term is significant and negative (p = .01), suggesting that participants who receive OFB are able to improve decision performance significantly more than those who do not receive OFB, providing support for H2.

To further explore H2, we separately evaluate the effect of iteration on both the group that received OFB and the group that did not receive OFB. Table 6 presents the results of a two-way, simplified ANOVA model for both groups. Consistent with Panel C of Table 4, we find that for participants who did not receive OFB, *Iteration* is not significant (p = .80). This result implies that without OFB, participants were unable to improve their profit predictions. Those who received OFB, however, were able to significantly improve their profit predictions (p = .001). Taken together, this suggests that while mere replication is not sufficient to lead to improved judgments, OFB enables significant improvement in a complex task.

Additional Analysis

Prior feedback papers have also used judgment achievement (r_a) as a proxy for judgment performance (Leung and Trotman, 2005; Hirst et al. 1999). Judgment achievement is a lens model parameter[14] that is calculated as the correlation between individuals' profit predictions and actual plant profits. A high achievement score is indicative of more accurate judgments. Table 7 reports our regression results with performance measured as judgment achievement. As shown in Table 7, only participants who

[14] See Luft and Shields (2001) for a description of the other lens model parameters. Luft and Shields investigate the other lens model parameters to explain why there was a difference between participants who received the expense treatment and participants who received the investment treatment. Because we find no fixation effect, there is no reason to investigate the additional lens model parameters in order to explain a fixation effect. Further, most of the prior OFB studies do not investigate the additional lens model parameters as there is likely no marginal contribution in this context (Bonner and Walker, 1994; Earley, 2003; Earley, 2001; Leung and Trotman, 2005; Leung and Trotman, 2008).

Table 6. Two-Way ANOVA for Different Values of *OFB*

MODEL: $APE = \beta_0 + \beta_1\, Expense + \beta_2\, Iteration + \varepsilon_i$

No Feedback (*OFB* = 0)	*F*-Value	*p*-value (two-tailed)
Intercept (β_0)	582.14	<0.01
Expense (β_1)	0.13	0.72
Iteration (β_2)	0.07	0.80

Feedback (*OFB* = 1)	*F*-Value	*p*-value (two-tailed)
Intercept (β_0)	453.32	<0.01
Expense (β_1)	0.48	0.49
Iteration (β_2)	13.78	<0.01

APE: the average absolute prediction error for each participant

Expense: equals 0 if intangible expenditures are labeled as investments and 1 if intangible expenditures are labeled as expenses

Iteration: equals 0 for the first task iteration and 1 for the second task iteration

OFB: equals 0 if the participant is not presented with outcome feedback during the second task iteration and 1 if the participant is presented with outcome feedback during the second task iteration

received OFB were able to improve performance ($p = 0.03$). Thus, our results remain unchanged when we use this alternative measure of performance.

V. Discussion and Conclusion

This paper examines the effect of task repetition and OFB on decision performance in a complex accounting task. We demonstrate that OFB is capable of improving performance in a relatively complex task. Prior psychology studies have repeatedly argued that OFB is ineffective and in certain cases even harmful when offered in complex tasks (Hammond and Summers, 1972; Todd and Hammond 1965; Hammond et al. 1973; Balzer et al., 1989). However, these prior psychology studies use abstract tasks in which the decision cues and criterion have no real-world application. Consequently, some have speculated that in more meaningful decision environments, perhaps the effect of OFB may be different (Libby, 1981; Hirst et al., 1999). As the primary contribution of this paper, we find that OFB can improve performance in a complex task when the task is placed in a more meaningful learning environment.

Prior research indicates that for OFB to be effective, the judge must be able to work backwards from the outcome (i.e., the result) and develop explanations of the environment (Bonner and Walker, 1994). Because the judge must reason backwards from the outcome to infer the underlying environmental relationships, it may appear questionable whether OFB can promote learning in a complex task. Yet, this study suggests that OFB can still encourage learning in a complex environment.

Like Hirst et al. (1999), this study offers support for the use of OFB over a wider range of task conditions. In prior OFB-based accounting research, decision cues were always related, to some extent, to the criterion. In this study, however, only one of the four decision cues—the lagged decision cue—was related to the criterion, thus adding an extra element of task complexity. Real-life decision environments oftentimes present individuals with multiples cues, some relevant and others irrelevant. The task used in

Table 7. Two-Way ANOVA for Different Values of *OFB* with Performance Measured as Judgment Achievement (r_a)

MODEL: $Achievement = \beta_0 + \beta_1\ Expense + \beta_2\ Iteration + \varepsilon_i$

No Feedback (*OFB* = 0)	*F*-Value	*p*-value (two-tailed)
Intercept (β_0)	111.99	<0.01
Expense (β_1)	0.80	0.38
Iteration (β_2)	1.36	0.25

Feedback (*OFB* = 1)	*F*-Value	*p*-value (two-tailed)
Intercept (β_0)	37.10	<0.01
Expense (β_1)	0.55	0.46
Iteration (β_2)	5.03	0.03

Achievement: an individual's judgment achievement (r_a) measured as the correlation of an individual's actual profit predictions with the correct plant profits.

Expense: equals 0 if intangible expenditures are labeled as investments and 1 if intangible expenditures are labeled as expenses

Iteration: equals 0 for the first task iteration and 1 for the second task iteration

OFB: equals 0 if the participant is not presented with outcome feedback during the second task iteration and 1 if the participant is presented with outcome feedback during the second task iteration

this study allows us to investigate the effectiveness of OFB in such contexts. Our findings indicate that OFB can lead to improvements in performance even when the decision cues are noisy.

Our results also indicate that repetition (without OFB) does not improve judgment performance in a complex accounting task. Prior experimental economics research finds that learning can occur in a no-feedback environment (Weber, 2003; Rapoport et al., 2002; Grether, 1980). However, this study, like several previous accounting studies, finds that repetition (without OFB) does not lead to improvements in judgment performance (Bonner and Walker, 1994; Leung and Trotman, 2005). Future research could investigate the effect of mere repetition in more simplistic accounting tasks, as it is likely that the task used in this study was too complex to detect any improvements in performance from mere repetition alone.

Finally, we fail to replicate the fixation effect found in L&S. We note, however, that while L&S use MBA students as participants, we use MSA students. We argue that the lack of fixation may result because MSA students potentially have a more in-depth understanding of accounting terminology, thus enabling them to see through accounting labels and avoid fixation. This finding is important as it suggests that L&S's fixation results may not obtain with individuals who have more in-depth accounting knowledge.

Archival accounting research is oftentimes subjected to multiple replications when prior models are extended. Behavioral accounting research, however, is rarely replicated. We believe that future behavioral accounting research could benefit from replications of important prior findings. We do not necessarily encourage pure replications in the sense that all experimental design choices remain constant, but rather, we encourage replications such as Bamber et al. (2000) which test the boundary conditions under which important prior findings may either persist or diminish.

In particular, we encourage future replications which extend important prior findings by adding a second iteration along with some form of feedback. If a single-shot experiment intends to represent a task that actually includes feedback and repeated task performance, then the findings from such single-shot experiments may have reduced external validity. For example, Sprinkle (2000) finds that incentives enhance performance in a multiperiod task and notes that many single-shot decision-making experiments fail to detect incentive effects because, unlike the real world, they do not offer feedback. Similar findings in other behavioral accounting research that use single-iteration experimental designs may also be limited in application. For example, individuals who at first do not understand the meaning of an unaudited letter to the shareholders (Hodge 2001), the problems related to a lack of independence (Hirst, et al. 1995), or the different formats of presentation (Maines and McDaniel 2000) may develop an understanding of these issues after making repeated judgments. Therefore, we encourage future research that extends important accounting studies which use single task experimental designs by assessing their respective results over multiple iterations.

References

Balzer, W. K., Doherty, M. E., and O'Connor Jr., R. (1989). Effects of Cognitive Feedback on Performance. *Psychological Bulletin* 106 (November): 410–433.

Bloomfield, R. and Libby, R. (1996). Market Reactions to Differentially Available Information in the Laboratory. *Journal of Accounting Research* 34 (Autumn): 183–207.

Bloomfield, R., Libby, R., and Nelson, M. (1999). Confidence and the Welfare of Less-Informed Investors. *Accounting, Organizations and Society* 24 (8): 623–647.

Bloomfield, R., Libby, R., and Nelson, M. (2003). Do Investors Over-rely on Old Elements of the Earnings Time Series? *Contemporary Accounting Research* 20 (Spring): 1–31.

Bloomfield, R. and Wilks, T. (2000). Disclosure Effects in the Laboratory: Liquidity, Depth, and the Cost of Capital. *The Accounting Review* 75 (January): 13–42.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley CA.

Butler, D. L. and Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research* 65: 245-281.

Cheung, Y. and Friedman, D. (1997). Individual Learning in Normal Form Games: Some Laboratory Results. *Games and Economic Behavior* 19: 46-76.

Domjan, M. (2005). Pavlovian Conditioning: A Functional Perspective. *Annual Review of Psychology* 56 (1): 179–206.

Earley, C. E. (2003). A Note on Self-Explanation as a Training Tool for Novice Auditors: The Effects of Outcome Feedback Timing and Level of Reasoning on Performance. *Behavioral Research in Accounting* 15: 111-123.

Elliott, W. B. (2006). Are Investors Influenced by Pro Forma Emphasis and Reconciliations in Earnings Announcements? *The Accounting Review* 81 (January): 113–133.

Elliott, W. B., Hodge, F. D., Kennedy, J. J., and Pronk, M. (2007). Are M.B.A. Students a Good Proxy for Nonprofessional Investors? *The Accounting Review* 82 (1): 139–168.

Epstein, B. J. (2007). Information Overload Can Threaten Sound Decision Making. *CPA Journal* 77 (3): 6–10.

Erev, I. and Roth, A. 1998. Predicting how People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review* 88 (4): 848-881.

Frederickson, J. and Miller, J. (2004). The Effects of Pro Forma Earnings Disclosures on Analysts' and Nonprofessional Investors' Equity Valuation Judgments. *The Accounting Review* 79 (July): 667–686.

Grether, D.D. 1980. Bayes Rule as a Descriptive Model: The Representativeness Heuristic. *The Quarterly Journal of Economics* 96 (3): 537-557.

Hammond, K. R. and Summers, D. A. (1972). Cognitive Control. *Psychological Review* 79: 58-67.

Hammond, K. R., Summers, D. A., and Deane, D. H. (1973). Negative Effects of Outcome-Feedback in Multiple-Cue Probability Learning. *Organizational Behavior and Human Performance* 9: 30–34.

Hirst, D., Jackson, K., and Koontz, L. (2003). Improving Financial Reports by Revealing the Accuracy of Prior Estimates. *Contemporary Accounting Research* 20 (Spring): 165–193.

Hirst, D., Koontz, L., and Miller, J. (1999). The Joint Effect of Management's Prior Forecast Accuracy and the Form of Its Financial Forecast on Investor Judgment. *Journal of Accounting Research* 37 (Supplement): 101–124.

Hirst, D., Koontz, L., and Simko, P. (1995). Investor Reactions to Financial Analysts' Research Reports. *Journal of Accounting Research* 33 (Autumn): 335–351.

Hirst, M. K. and Luckett, P. F. (1992). The Relative Effectiveness of Different Types of Feedback in Performance Evaluation. *Behavioral Research in Accounting* 4: 1–22.

Hodge, F. (2001). Hyperlinking Unaudited Information to Audited Financial Statements: Effects on Investor Judgments. *The Accounting Review* 76 (October): 675–691.

Hodge, F., Kennedy, J., and Maines, L. (2004). Does Search-Facilitating Technology Improve the Transparency of Financial Reporting? *The Accounting Review* 79 (July): 687–703.

Kennedy, J., Mitchell, T., and Sefcik, S. (1998). Disclosure of Contingent Environmental Liabilities: Some Unintended Consequences. *Journal of Accounting Research* (Autumn): 257–277.

Krische, S. (2005). Investors' Evaluations of Strategic Prior-Period Benchmark Disclosures in Earnings Announcements. *The Accounting Review* 80 (January): 243–268.

Leung, P. W. and Trotman, K. T. (2005). The Effects of Feedback Type on Auditor Judgment Performance for Configural and Non-Configural Tasks. *Accounting, Organizations and Society* 30: 537-553.

Libby, R. *Accounting and Human Information Processing: Theory and Applications.* Prentice Hall: 1981.

Lipe, M. (1998). Individual Investors' Risk Judgments and Investment Decisions: The Impact of Accounting and Market Data. *Accounting, Organizations and Society* 23 (7): 625–640.

Luft, J. and Shields, M. D. (2001). Why Does Fixation Persist? Experimental Evidence on the Judgment Performance Effects of Expensing Intangibles. *The Accounting Review* 76 (4): 561–587.

Luft, J. and Shields, M. D. (2009). Psychology Models of Management Accounting. Working Paper. Michigan State University.

Maines, L. and Hand, J. (1996). Individuals' Perceptions and Misperceptions of Time Series Properties of Quarterly Earnings. *The Accounting Review* 71 (July): 317–336.

Maines, L. and McDaniel, L. (2000). Effects of Comprehensive Income Characteristics on Nonprofessional Investors' Judgments: The Role of Financial Statement Presentation Format. *The Accounting Review* 75 (April): 179–207.

Mercer, M. (2005). The Fleeting Effects of Disclosure Forthcomingness on Management's Reporting Credibility. *The Accounting Review* (April): 723–744.

Nelson, M., Krische, S., and Bloomfield, R. (2003). Confidence and Investors' Reliance on Disciplined Trading Strategies. *Journal of Accounting Research* 41 (June): 503–523.

Rapoport, A., D. Seale, and J. Parco. 2002. Coordination in the Aggregate Without Common Knowledge or Outcome Information. In R. Zwick and A. Rapoport (Eds.), *Experimental Business Research*. Norwell, MA: Kluwer Academic Press.

Shepherd, D. A. and Zacharakis, A. (2002). Venture Capitalists' Expertise: A Call for Research into Decision Aids and Cognitive Feedback. *Journal of Business Venturing* 17: 1-20.

Staddon, J. E. R. and Cerutti, D. T. (2003). Operant Conditioning. *Annual Review of Psychology*, 54: 115–144.

Staddon, J. E. R. (2003). Adaptive Behavior Learning. Cambridge. Cambridge University Press.

Todd, F. J. and Hammond, K. R. (1965). Differential Feedback in Two Multiple-Cue Probability Learning Tasks. *Behavioral Science* 10 (4): 429–435.

Weber, R.A. 2003. Learning with No Feedback in a Competitive Guessing Game. *Games and Economic Behavior* 44: 134-144.