This is to certify that the

dissertation entitled

EFFECTS OF MULTIPLE PERFORMANCE MEASURES, MULTICOLLINEARITY, AND TASK STRUCTURE ON INDIVIDUALS' JUDGMENT PERFORMANCE

presented by

Anne Magner Farrell

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Accounting and Information Systems

Michael D. Shields
Major professor

Date November 15, 2002

EFFECTS OF MULTIPLE PERFORMANCE MEASURES, MULTICOLLINEARITY, AND TASK STRUCTURE ON INDIVIDUALS' JUDGMENT PERFORMANCE

By

Anne Magner Farrell

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Accounting and Information Systems

2002

ABSTRACT

EFFECTS OF MULTIPLE PERFORMANCE MEASURES, MULTICOLLINEARITY, AND TASK STRUCTURE ON INDIVIDUALS' JUDGMENT PERFORMANCE

By

Anne Magner Farrell

This dissertation empirically investigates how the use of multiple performance measures affects individuals' judgment performance. Specifically, it provides theory-based experimental evidence on how the number of performance measures used to measure a particular organizational objective and the multicollinearity in those measures interactively affect individual judgment performance in a prediction task. Further, it investigates how a change in the structure of this task affects judgment performance. Measures of judgment performance capture how accurately individuals estimate the relations between and among the performance measures, and how consistently they apply the relations they estimate to make predictive judgments.

Results suggest that judgment performance is an interactive function of the number of accounting measures and their multicollinearity, but task structure has no effect on judgment performance. An increase in the number of measures results in less accurate estimates of the relations between performance measures and less consistent application of those estimates when multicollinearity is high but not when it is low. Supplementary analyses suggest that individuals know multicollinearity is important to estimates of relations between performance measures but do not know how to incorporate it into their judgments. This dissertation concludes by identifying its contributions, limitations, and possible directions for future research.

Dedicated to my husband and best friend, Tony,

to my wonderful children, Abby and Max,

and to my loving and supportive parents, Marie and Charlie Magner.

ACKNOWLEDGEMENTS

Completion of this degree would not have been possible without the incredible support of my committee: Michael Shields (chair), Joan Luft, Ranjani Krishnan, and Daniel Ilgen. Throughout my program of studies, they cheered me on in my brightest moments and provided advice and support through the darkest ones. Special thanks go to Mike and Joan for allowing me to use in this dissertation a data generation program they had developed for another project.

Much gratitude is also due to all of the other faculty members in the Department of Accounting and Information Systems who create an environment in which Ph.D. students can flourish. I am lucky to have had the chance to work with all of them.

Finally, I am especially thankful for the love and support (and tolerance!) from my husband, Tony, my children, Abby and Max, my parents, Charlie and Marie Magner, my siblings, Mary, Mike, and Pat, and their families, all of my in-laws, and the many friends near and far who have always been there for me.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: THEORETICAL DEVELOPMENT AND HYPOTHESES 8
- Subjective Judgments in Organizations 8
- Relations Between Performance Measures 9
- Effects of Number of Measures, Multicollinearity, and Task Structure on Judgment Performance 11
  - Processing Error and Number of Cues 15
  - Attentional Error and Multicollinearity 16
  - Attentional-by-Processing Error and Multicollinearity 17
  - Total Judgment Error 18
  - Total Judgment Error and Task Structure 20
- Dimensions of Judgment Performance 23

CHAPTER 3: EXPERIMENTAL DESIGN 26
- Participants and Power Analysis 26
- Independent Variables 26
- Experimental Setting 27
- Procedures 34
- Dependent Variables 36

CHAPTER 4: RESULTS OF EXPERIMENT 40
- Analysis of Heuristics Used 40
- Tests of Randomization and Sensitivity of Results 41
- Tests of Hypotheses and Supplemental Analyses 42
  - Test of H1: Accuracy of Estimated Cue-Criterion Weights 42
  - Discussion of Results for H1: Accuracy of Estimated Cue-Criterion Weights 43
  - Test of H2: Judgment Consistency 46
  - Discussion of Results for H2: Judgment Consistency 47

CHAPTER 5: DISCUSSION AND CONCLUSION 49
- Synthesis of Results 49
- Limitations 52
- Contributions 55
- Implications for Practice 58
- Possible Directions for Further Research 59

ENDNOTES....78

APPENDICES....81
Appendix A: Small-Scale Empirical Investigation of Heuristics Used In Predictive Judgment Task....82
Appendix B: Illustration of Computation of Regression Weights for Models With Two or Five Independent Variables....87
Appendix C: Notes on Effects of Multicollinearity on Estimated Weights In OLS Regression....96
Appendix D: Experimental Materials....100

BIBLIOGRAPHY....126

LIST OF TABLES

TABLE 1: PARAMETERS FOR EXPERIMENTAL DATA SETS.....63

TABLE 2: SAMPLE OF LEARNING DATA PROVIDED TO PARTICIPANTS.....66

TABLE 3: DESCRIPTIVE STATISTICS.....67

TABLE 4: HYPOTHESIS 1 RESULTS -- ACCURACY OF ESTIMATED CUE-CRITERION WEIGHTS.....68

TABLE 5: HYPOTHESIS 2 RESULTS -- JUDGMENT CONSISTENCY.....70

TABLE A1: OCCURENCES OF NEGATIVE REGRESSION WEIGHTS IN DATA SETS WITH HIGH MULTICOLLINEARITY.....88

TABLE A2: RELATIVE WEIGHTS OF r_{yi} IN COMPUTING b_i WITH DIFFERING DEGREES OF MULTICOLLINEARITY.....95

LIST OF FIGURES

FIGURE 1: RELATIONS BETWEEN MULTIPLE PERFORMANCE MEASURES....72

FIGURE 2: SOURCES OF ERROR WITH MULTIPLE MEASURES, MULTICOLLINEARITY AND TASK STRUCTURE....73

FIGURE 3: EXPECTED FORM OF EFFECTS OF NUMBER OF CUES, MULTICOLLINEARITY, AND TASK STRUCTURE ON JUDGMENT PERFORMANCE....74

FIGURE 4: DIAGRAM OF PERFORMANCE MEASUREMENT SYSTEM PROVIDED TO PARTICIPANTS....75

FIGURE 5: HYPOTHESIS 1 RESULTS -- ACCURACY OF ESTIMATED CUE-CRITERION WEIGHTS....76

FIGURE 6: HYPOTHESIS 2 RESULTS -- JUDGMENT CONSISTENCY....77

CHAPTER 1: INTRODUCTION

Organizations increasingly use multiple performance measures instead of a single measure to provide information about important objectives, particularly more difficult-to-quantify objectives like learning and growth, innovation, quality, and employee or customer satisfaction. This trend has been in part driven by the popularity of strategic performance measurement systems like Kaplan and Norton's "balanced scorecard", which recommend that organizations link objectives and their chosen performance measures together in a cause-and-effect chain. Purported benefits of these performance measurement systems are that employees have clearer action-to-performance links, and that using multiple performance measures for a given organizational objective can reduce noise in the measurement of that objective (Kaplan and Norton 1992, 1993, 1996a-c, 2000, 2001; Balkcom, Ittner and Larcker 1997; Lambert 1998; Sjoblom 1998; Stivers et al. 1998; Kaplan and Tempest 1999; Hertenstein and Platt 2000).

Although using multiple performance measures does provide important incremental information to individuals in the organization, there are concerns that "...a large number of measures can reduce performance by exceeding managers' [cognitive] processing capabilities when making judgments..." (Ittner and Larcker 1998, p. 226). If this reduction in individual judgment performance occurs, then performance at the organizational level is also reduced since resources used to collect this information will be wasted and the benefits of having more information will not be realized (Stivers et al. 1998). When addressing the question of how many measures are too many, practitioner literature often implicitly assumes that as long as the number of measures is kept below

some threshold of cognitive overload, judgment performance is not affected (Kaplan and Norton 1992, 1993, 1996a-c, 2000, 2001; Simons and Davila 1998). This dissertation, however, predicts that the effect of the number of measures on judgment performance depends on characteristics of the measures (e.g., multicollinearity) and the structure of the judgment task in which the measures are used, even below that implicit threshold of cognitive overload.

Specifically, in prior literature that investigated judgment performance with multiple measures and with multicollinearity, individuals were given very large, abstract data sets (often with 200 observations or more) to learn relations in data and to make subsequent judgments. In practice, however, individuals who use organizational data to learn relations between performance measures and to make subsequent judgments frequently have very few observations (sometimes 10 or fewer). Further, prior literature suggests that individuals fail to incorporate multicollinearity into their judgments, but use of multiple performance measures for organizational objectives gives rise to the fact that individuals in practice do need to recognize and process multicollinear measures. A task structure that addresses the need to recognize multicollinearity is expected to result in increased judgment performance.

Suppose that the task structure is such that individuals' attention is focused on the *measures* of organizational objectives. If multiple performance measures are used to reduce noise in the measurement of a single organizational objective (i.e., multicollinearity is high), then an increase in the number of measures is expected to reduce judgment performance more than if those multiple measures are used to capture independent dimensions of the organizational objective (i.e., multicollinearity is low),

holding the predictive ability of the measures constant. Alternately, if the task structure focuses individuals' attention on the organizational objective *underlying the measures*, then it is expected that judgment performance will be reduced less by the negative interactive effect of increases in the number of measures and their multicollinearity. This dissertation provides theory-based experimental evidence of judgment performance with respect to the accuracy of individuals' estimates of the relations between causally-related performance measures (where OLS weights from a regression model are the standard for the most accurate estimate), and the consistency with which they apply those estimates in a predictive judgment task.

The task in this dissertation is one in which individuals have a set of past observations of performance measures for two causally-related organizational objectives, product quality and customer satisfaction (analogous to having a series of observations of independent and dependent variables). These past observations can be used to estimate relations among the objectives and the measures. The individuals then receive a series of potential values of the measures of one of the organizational objectives (e.g., a series of potential values for the product quality measures, which are the independent variables), and are asked to make predictive judgments about the values of the causally-related measures of the objectives (e.g., predictive judgments about measures of customer satisfaction, which are the dependent variables). This task is similar to how managers prepare budgets or analyses used in resource allocation decisions.

Prior research suggests that the total judgment error in this task can be decomposed into three types, and these errors decrease judgment performance. First, increasing the number of measures requires more cognitive processing, resulting in

processing errors, holding the predictive ability of the set of measures constant (Huber 1985; Wood 1986; Lee and Yates 1992; Bonner 1994). Second, other research finds that individuals frequently fail to incorporate the effects of multicollinearity when making judgments, thus committing *attentional errors* (Armelius and Armelius 1974; Brehmer 1974b; Lindell and Stewart 1974; Schmitt and Dudycha 1975; Libby 1981; Schum and Martin 1982; Klayman 1988; Maines 1990, 1996). When multicollinearity is low, attentional errors are less significant, because there is relatively little difference between the OLS weights from a regression analysis and estimates of the weights that fail to incorporate multicollinearity. However, as multicollinearity increases, attentional error increases since the difference between the OLS weights and weights estimated without adjustments for multicollinearity increases. Third, if individuals attempt to make predictions using inaccurate weights for the independent variables and see that their predictions are not close to observed values of the dependent variable because of attentional error, then further processing errors can arise when individuals attempt to make adjustments to their weights or judgments and such adjustments are imperfect (i.e., there is an interactive effect of *attentional errors and processing errors*).

Prior research also suggests that task structure can affect judgment performance by influencing the difficulty of cognitive processing and focusing attention on different parts of the task (Simon 1978; Getzels 1982; Schum and Martin 1982; Trabasso 1982; Payne, Bettman and Johnson 1992; Goodwin and Wright 1993, 1994; Messier 1995; Ruscio 2000). This dissertation investigates whether the structure of the predictive judgment task described above interacts with the number of measures and their multicollinearity to affect total judgment error. The task structure can be such that

attention is either focused on *only* the measures for organizational objectives (hereafter called an *indicator structure*), or on *both* the measures *and* the organizational objectives (a *construct structure*). The construct structure decomposes relations in the task into two types, consistent with relations in structural equation modeling (SEM) – relations among a set of performance measures for a given organizational objective, and relations between those performance measures and measures of another organizational objective to which they are causally linked. I predict that a construct structure will reduce the effects of both processing errors (by decomposing cognitive processing requirements into smaller parts) and attentional errors and attentional-by-processing errors (by focusing individuals' attention on multicollinearity).

Consistent with a construct structure, some organizations that use multiple performance measures ask individuals to provide a single summary rating for each strategic objective based on multiple measures of that objective before they make a summary judgment about overall performance across all strategic objectives (Ittner, Larcker and Meyer 2002). Consistent with an indicator structure, other organizations do not ask individuals to make summary ratings for each strategic objective, but instead have them use the individual measures to make the judgment of overall performance (Ernst & Young, 2002).

This dissertation contributes to the scholarly literature in accounting and psychology in three ways. First, while performance measurement systems that map key organizational objectives in a cause-and-effect chain are increasingly popular in practice (Kaplan and Norton 1992, 1993, 1996a-c, 2000, 2001), there is limited research on the effects of the design of these performance measurement systems on individual judgment

performance (see Sprinkle 2002; exceptions include Krumwiede, Eaton and Swain 2000; Lipe and Salterio 2000, 2002; Ullrich and Tuttle 2000; Luft and Shields 2001). The existing research does not focus on how the number of measures and multicollinearity interact to affect judgment performance, holding the predictive ability of the set of measures constant, and how differences in task structure may affect judgment performance. However, many cause-and-effect performance measurement systems may be designed with links that have varying numbers of measures and levels of multicollinearity, so it is important to predict and explain how individual judgment performance may differ at these various links because of these design factors. Such differences, and whether changes in task structure can reduce them, are of interest to individuals in organizations in which predictive judgments using these performance measures are the basis for resource allocation decisions, and to designers of performance measurement systems.

Second, cause-and-effect performance measurement systems which use multiple measures for each organizational objective strongly resemble structural equation models. However, I found no prior research that examines whether decomposing a judgment task into parts that resemble those of structural equation models results in different judgment performance than in a non-decomposed task. Third, much of the prior research that examines individual judgments based on multicollinear data was conducted with abstract tasks, and the results are difficult to interpret because the dependent variables were correlational measures that were inflated with multicollinearity (Naylor and Schenck 1968; Armelius and Armelius 1974; Brehmer 1974b; Lindell and Stewart 1974; Schmitt and Dudycha 1975; Libby 1981, p. 42; Ashton 1982, p. 37). This dissertation uses a

more concrete business judgment task, and the dependent variables are less prone to this interpretation problem and thus are useful when judgment performance with multicollinear data is of interest.

The remainder of this dissertation is organized as follows. Chapter 2 begins with a discussion of the use of subjective judgments in organizations and of relations between performance measures, and proceeds with a review of the prior literature and the development of hypotheses about how the number of performance measures, their multicollinearity, and task structure interact to affect judgment performance. Chapter 3 describes the experimental design, and Chapter 4 presents results of the experiment. Chapter 5 is a synthesis of the dissertation and the results, its limitations and contributions, and possible directions for further research.

CHAPTER 2: THEORETICAL DEVELOPMENT AND HYPOTHESES

Subjective Judgments in Organizations

While it may seem that performance measures should be selected because of their informativeness (i.e., statistical predictive ability) about important organizational objectives (see, for example, Holmstrom 1979), often they are selected on the basis of management intuition about this informativeness. For example, one financial services company used subjectively-developed cause-and-effect models to choose their performance measures (Simons and Davila 1998), while a hotel chain chose their key drivers of performance through management discussion and consensus (Banker, Potter and Srinavasan 2000).

Managers in organizations frequently do not use statistical models to guide their choice and use of performance measures because they may not have the resources needed to develop or use the models, the models may assume particular conditions while management believes the organization is operating an environment with different conditions, or the data needed to estimate models may be costly to obtain or to adjust for the effects of unusual events. Further, employees may not use performance measures as inputs into statistical models to aid their judgments because they may not be given access to or may not understand the underlying models, may be skeptical of their output, may believe that using models results in a loss of control, or may believe they can outperform the model (Goodwin and Wright 1994; Kaplan and Norton 1996b, 2001). Because of a reluctance to use statistical models, it is important to investigate how and how well

individuals make subjective judgments, and what factors affect that judgment performance.

Relations Between Performance Measures

When multiple performance measures are used to measure organizational objectives, four types of relations are important. These relations can be described by reference to SEM. SEM provides statistical estimates of relations between unobservable variables (called *constructs*, which in this dissertation are organizational objectives such as innovation, learning, employee or customer satisfaction, or quality) and observable measures of them (called *indicators*, which in this dissertation are performance measures). While indicators are directly measured, constructs are not directly measured but can be estimated by SEM based on the correlations among the indicators (e.g., factor analyses). SEM provides simultaneous estimates of the relations between all constructs and all indicators in a measurement system. However, the relations in the system can also be decomposed, and that change in focus can result in the use of different tools to estimate those relations.

For example, assume that two causally-related objectives, X and Y (constructs), are measured by their respective performance measures (indicators), $\{x_1 ... x_n\}$ and $\{y_1 ... y_m\}$ (Figure 1). For example, many organizations are interested in how product quality (X) affects customer satisfaction (Y), but since product quality and customer satisfaction are themselves unobservable, multiple measures are used to proxy for each. The causal relation of most interest to organizations is the relation between the organizational objectives themselves. This is illustrated by the bold line from X to Y (relation 1) in Figure 1. In SEM, relations between constructs are a function of the

relations between their chosen indicators, so this dissertation does not investigate how well individuals estimate relations between the organizational objectives directly. Instead, this dissertation investigates how well individuals estimate the remaining three types of relations and use those estimates to make predictive judgments (Figure 1):

- The relations between the performance measures of the causally-related objectives (relation 2, the dashed lines from each of the measures $x_1 ....x_n$ to each of the measures $y_1 ...y_m$); these relations can be estimated by regression analysis.
- The relations between each performance measure and the objective it measures (relation 3, the solid lines from an objective to each of its measures); these relations can be estimated by the component scores from factor analysis.
- The correlations between performance measures (multicollinearity) for an objective (relation 4, the dots between the measures $x_1 ....x_n$ of objective *X* and $y_1 ...y_m$ of objective *Y*); the magnitude of these relations can affect the weights on independent variables in a regression analysis (Relation 2) and the component scores in a factor analysis (Relation 3).

INSERT FIGURE 1

Relation 2 in Figure 1 is relevant when individuals want to make predictions from one set of causally-related performance measures to another, such as when they are preparing budgets or analyses used for resource allocation decisions. The first task structure investigated in this dissertation, the *indicator structure*, involves making predictive judgments for this type of task. Relation 3 in Figure 1 is relevant when

individuals attempt to understand how well their chosen performance measures proxy for an unobservable organizational objective, want to estimate the value of an organizational objective based on its set of performance measures, or make resource allocation decisions based on the level of an organizational objective or directed at changing that level. The second task structure investigated in this dissertation, the *construct structure*, expands the indicator structure to incorporate estimates of the underlying value of the organizational objective into the predictive judgment task. Note that regardless of whether individuals are interested in estimating relation 2 or relation 3, relation 4 should be considered.

Because the purpose of this dissertation is to investigate how the number of performance measures and their multicollinearity affect judgment performance independent of other factors, the statistical predictive ability of the set of casually-related measures (i.e., Figure 1, relation 2) is held constant across different numbers of measures and their multicollinearity. It would not be surprising if judgment performance were better when the predictive ability of measures was higher, so investigation of this issue would not be interesting. An investigation of differences in judgment performance holding predictive ability constant, however, is interesting because it can provide insight into whether the inclusion of more measures that may be highly correlated with other measures is worth the cost.

Effects of Number of Measures, Multicollinearity, and Task Structure on Judgment Performance

Suppose an individual is interested in the relations between the performance measures of two causally-linked objectives (i.e., the focus is on relation 2 in Figure 1). Hereafter, consistent with psychology and accounting literature on judgment and decision making, performance measures $x_1 ... x_n$ will be referred to as *cues*, analogous to

independent variables in a regression model, and performance measure y_i will be referred to as the *criterion*, analogous to the dependent variable in a regression (see Libby 1981, pp. 18-21; Ashton 1982, pp. 14-18). Consistent with the indicator structure, assume that the individual will use a cross-sectional set of *past* observations of the cues and criterion to estimate the weights for the cues, and will then apply those weights to a series of potential values of the cues to make predictive judgments of the criterion. Because the individual's goal is to use the cue values to make predictive judgments about the criterion, attention is directed to the measures of organizational objectives (i.e., the indicators) and not on the underlying organizational objectives themselves.

To complete the judgment task, individuals first use the past observations of the cues and criterion to estimate the weights they will place on the cues. Prior research that examines the heuristics individuals use to determine the cue-criterion weights in this context is limited. However, that prior research coupled with a small-scale empirical investigation of how individuals do this task (Appendix A) indicates that there are two primary heuristics that individuals use – a difference heuristic (Hutchinson and Alba 1997) and an equal-weight heuristic (Peterson, Hammond and Summers 1965; Brehmer 1973a; Nisbett, Zukier and Lemley 1981; Bloomfield, Libby and Nelson 1998a, b). As can be seen in the following descriptions of how these heuristics would be used to determine cue-criterion weights for this task, an important feature of both heuristics is that individuals tend to focus on bivariate cue-criterion relations, not on partial correlations or multiple regression weights; there is no evidence that individuals explicitly try to incorporate multicollinearity into their estimation of the bivariate cue-criterion weights when there are multiple cues.

With a difference heuristic, each cue-criterion weight is estimated by comparing the change in the criterion to a corresponding change in the selected cue using pairs of the past cue-criterion observations. Individuals may use only one pair of cue-criterion observations (e.g., the observations with the lowest and highest values for a cue) or may use multiple pairs of observations and combine the results in some way (e.g., use the mean or median). This process is repeated until a weight has been estimated for each cue.

With an equal-weight heuristic, the same weight is applied to each cue regardless of the actual relations in the task. The weight may be determined by applying a difference heuristic to one cue and using that as the weight for all cues, or it may be based on the inverse of the number of cues in the task (i.e., 1/number of cues), the latter of which was the case in the small-scale empirical investigation (Appendix A). In general, this heuristic is most applicable when the cues and criterion have compatible scales, such as when they are all in dollars or percentage variation from budget (Tversky, Sattath and Slovic 1988; Slovic, Griffin and Tversky 1990).

Although it is not clear which of these two heuristics an individual might choose to use to estimate cue-criterion weights, there is no evidence that the number of cues affects their choice (Payne 1976; Payne, Bettman and Johnson 1990; Bonner 1994; see also Appendix A). Therefore, it is assumed that these are the two heuristics individuals will apply to this task. However, prior research indicates that within a task, an individual sometimes will switch between these heuristics when making multiple predictions (Payne 1976; Bettman, Johnson and Payne 1990; Bonner 1994; see also Appendix A). The following explains why within-task switching can occur.

After the cue-criterion weights are estimated using either heuristic, individuals can take one of two approaches to complete the judgment task (Appendix A). First, individuals can apply the estimated weights to the potential set of cue values to make predictive judgments of the criterion, without any check on the accuracy of their weights or judgments. Alternatively, individuals can check the accuracy of their heuristic by applying the estimated weights to a set of cue values from the past cue-criterion observations they used to estimate the weights to compute a judgment for the criterion, and compare this judgment to the observed value of the criterion from those observations. If they find that the resulting judgments are inaccurate, then they can make imperfect adjustments (e.g., change the values of the weights or judgments by some constant or percentage), or switch to another heuristic to reestimate weights and repeat the process (see Appendix A for evidence on the adjustment process). Once they determine their estimated weights or adjustments are satisfactory, they apply them to the set of potential cue values to make their predictive judgments. Therefore, individuals tend to follow either a two-step cognitive process (estimate cue-criterion weights using one of the two heuristics and then apply the weights to the judgment task) or a three-step cognitive process (estimate cue-criterion weights using one of the two heuristics, check the accuracy of the weights and reestimate the weights or adjust if deemed necessary, and then apply the weights to the judgment task).

Use of these heuristics and cognitive processes in this task can result in three types of judgment error. The first is *processing error*, or mathematical errors that arise from mental computations done when estimating the cue-criterion weights and applying those weights to the cues. The second is *attentional error*, or error that results from

ignoring multicollinearity when estimating cue-criterion weights. The third is *attentional-by-processing error*, or error that results from ignoring multicollinearity and then during cognitive processing attempting to adjust for the effects of the attentional error. The number of measures and their multicollinearity affect the magnitude of *total judgment error* (i.e., processing error plus attentional error plus attentional-by-processing error), and in turn affect the *accuracy* of an individual's cue-criterion weights (i.e., how close the estimated weights are to statistically-estimated cue-criterion weights), and the *consistency* with which he or she applies those weights when making predictive judgments (i.e., how invariant an individual is in applying his or her estimated weights).

Estimates of an individual's cue-criterion weights are based on an OLS regression of his or her predictive judgments of the criterion on the set of potential cue values (i.e., an individual's policy-capturing model; see Libby 1981, p. 20 and Ashton 1982, p. 16). Statistically-estimated cue-criterion weights are based on an OLS regression of the past observations of the criterion on the corresponding cues (i.e., the environmental model; see Libby 1981, p. 20 and Ashton 1982, p. 16). The expected effect of the number of cues, their multicollinearity and task structure on these errors and thus on judgment accuracy and consistency are discussed in the following sections.

Processing Error and Number of Cues

Suppose that the number of cues in a task is two or five. Use of two cues is relevant since this dissertation is concerned with multiple performance measures and thus two is the lowest possible level; use of five is relevant since a review of practitioner literature on multiple-measure performance measurement systems indicates that the most

measures used to proxy for an organizational objective was five (Kaplan and Norton 1993, 1996a, 2000, 2001; Kaplan and Tempest 1999).

Prior research suggests that as the number of cues in a task increases, the number of cognitive mathematical operations that must be completed when applying one of the two heuristics to the task increases. Because there is a chance of processing error in each of these operations, the magnitude of the expected total processing error increases and thus expected judgment performance decreases (Huber 1985; Wood 1986; Lee and Yates 1992; Bonner 1994).[1]

This expected increase in processing error will occur if individuals use either the difference or the equal-weight heuristic, since increases in the number of cues requires computation of a larger number of cue-criterion weights (for the difference heuristic only), application of those weights to a larger number of cue values, and estimation of a larger number of adjustments to weights or judgments, if applicable. Because of more processing errors, estimates of cue-criterion weights will be farther from the OLS weights for the task. Further, if individuals check the accuracy of their heuristic and find it insufficiently accurate because their estimated weights and judgments include more processing error, then they may try to "hedge their bets" by switching their heuristic or imperfectly adjusting their weights or judgments throughout the task, resulting in lower judgment consistency (Payne 1976; Huber 1985; Payne, Bettman and Johnson 1988, 1990; Bonner 1994).

Attentional Error and Multicollinearity

Suppose further that in the task examined here, multicollinearity in the cues is low or high. Prior research using different types of tasks indicates that individuals do not

typically incorporate multicollinearity into their judgments, which then results in attentional error (Armelius and Armelius 1974; Brehmer 1974b; Lindell and Stewart 1974; Schmitt and Dudycha 1975; Libby 1981; Schum and Martin 1982; Klayman 1988; Maines 1990, 1996).[2] Other research suggests that when estimating weights to be placed on multiple cues, individuals focus on bivariate cue-criterion relations and not on incorporating cue-cue relations into those estimates (Armelius and Armelius 1974; Hutchinson and Alba 1997). Recall that with both the difference and equal-weight heuristics, it appears that individuals focus their attention on estimating bivariate cue-criterion relations, which indicates that individuals in this setting will not incorporate multicollinearity (see Appendix A for further evidence).

Regardless of whether the difference or the equal-weight heuristic is used, when multicollinearity is present ignoring it results in attentional error, leading to decreases in judgment performance.[3] If multicollinearity is low and individuals ignore it, then attentional error is less significant since an individual's estimates of cue-criterion weights will more closely approximate the cue-criterion Pearson correlations (at the extreme of no multicollinearity, bivariate cue-criterion Pearson correlations are equal to the partial correlations and the OLS weights from a multiple regression; see Appendix B). If multicollinearity is high and individuals ignore it, however, then attentional error results in estimated cue-criterion weights that are farther from the OLS weights.

Attentional-by-Processing Error and Multicollinearity

If multicollinearity is high and individuals ignore it when making their subjective estimate of the weight for each cue-criterion relationship, then their resulting weights will be too high since redundant information in the cues will in essence be double-counted.[4]

If individuals proceed to check the accuracy of their heuristics by applying their estimated cue-criterion weights to a set of cue values and comparing that to the observed value of the criterion, then large differences in the predicted and observed criterion values will result because of this double-counting. These large differences may prompt them to make imperfect adjustments to their weights or judgments (e.g., they may lower the estimated weights for some cues and not others, or they may lower all the weights by the same amount or percentage), or to switch heuristics and repeat the estimating and checking process. This increases cognitive processing requirements and thus processing error. Therefore, attentional error and any subsequent processing error it may generate result in subjective estimated cue-criterion weights that are farther from the OLS weights for the task. Further, when high multicollinearity is ignored, because the cue-criterion weights estimated using either heuristic result in preliminary judgments that are farther from past values of the criterion, individuals may be less certain their heuristics are effective, prompting them to switch heuristics or make imperfect adjustments to weights or judgments, which results in lower judgment consistency (Payne 1976; Huber 1985; Payne et al. 1988, 1990; Bonner 1994).

Total Judgment Error

For the indicator structure, the graph on the left side of Figure 2 illustrates how the number of cues and their multicollinearity ordinally interact to increase total judgment error. First, assume that there are two cues. If multicollinearity is low, then there is processing error but no attentional error, since ignoring multicollinearity is not detrimental when estimating cue-criterion weights. If multicollinearity is high and individuals do not attend to it, then they are ignoring one cue-cue relation, causing

attentional error because two cue-criterion weights are farther from the OLS weights than they would be if there were low multicollinearity. Any adjustment made to cue-criterion weights or judgments gives rise to more cognitive processing requirements, thus increasing attentional-by-processing error.

INSERT FIGURE 2

Now assume there are five cues. If multicollinearity is low, then processing error increases over that for two cues because there are more cue-criterion weights to be estimated and more cue values to which those weights must be applied to formulate a judgment. If multicollinearity is high and individuals do not attend to it, then they are ignoring ten cue-cue relations ([n (n-1) / 2], where n is the number of cues), causing attentional error because five cue-criterion weights are farther from the OLS weights than they would be if there were low multicollinearity. If imperfect adjustments are made to weights or judgments, then there are more cognitive processing requirements than in the two-cue case, and thus greater attentional-by-processing error.

Increases in total error will result in decreases in judgment performance. Estimated cue-criterion weights will be farther from the OLS weights due to higher total error. Increases in total error will prompt individuals to be less certain of the effectiveness of their heuristic, leading them to change their heuristic or make imperfect adjustments to their weights or judgments, which will result in lower judgment consistency.

Total Judgment Error and Task Structure

Prior literature finds that task structure can affect judgment performance by influencing the difficulty of cognitive processing and focusing attention on different parts of the task (Simon 1978; Getzels 1982; Trabasso 1982; Payne et al. 1992). Recall that the discussion in the prior sections was based on the indicator structure, in which individuals' attention is focused on the measures of organizational objectives (i.e., the indicators) and not on the underlying organizational objectives themselves. Alternatively, the construct structure decomposes the relations in the task into two types, consistent with the relations in SEM – the cue-cue relations (i.e., relation 4 in Figure 1), and the cue-criterion relations (i.e., relation 2 in Figure 1). This task structure focuses individuals' attention on the underlying values of organizational objectives (i.e., the constructs) as well as on the measures of the objectives (i.e., the indicators). In a different task, Schum and Martin (1982) found that individuals incorporate multicollinearity into their judgments more often when using a decomposition approach to a task. Goodwin and Wright (1993, 1994) and Messier (1995) propose that any decomposition of a task that draws attention to its structure should improve judgment performance, and Ruscio (2000, p. 146) found that stimulation of "effortful cognitive processes" improves judgment performance by helping individuals better estimate the validity of measures in a set.[5]

With a construct structure, individuals would first estimate the underlying value of an objective based on the series of past cue values (i.e., estimates of relation 3 in Figure 1), focusing attention on multicollinearity (i.e., relation 4 in Figure 1). Estimating the underlying value of the objective based on its measures is critical when making

judgments about the level of the objective or directed toward influencing that level; an added benefit is that this may also lead individuals to consider whether the measures are valid and reliable proxies for the objective, or whether a change in measures is warranted. Next, individuals would use the past cue-criterion observations to estimate cue-criterion weights, and make judgments of the criterion based on the series of potential cue values (as is done with an indicator structure), focusing attention on the cue-criterion relations (i.e., relation 2 in Figure 1).

A construct structure is expected to result in lower total judgment error as compared to the indicator structure and thus higher accuracy of estimated cue-criterion weights. If a construct structure focuses individuals' attention on multicollinearity and they estimate that it is low, then they may conclude that each cue captures a different dimension of the underlying objective, or that at least some cues measure something other than the objective. Regardless of their conclusion, they are expected to more carefully estimate the cue-criterion relations that will ultimately be used as weights. This does not reduce the amount of cognitive processing required, but it may lead individuals to more carefully estimate the cue-criterion weights so that the only processing error that arises is from the application of those weights to the set of potential cue values (which is expected to be less significant than errors in estimation of the weights themselves). If a construct structure focuses individuals' attention on multicollinearity and they estimate that it is high, then they are expected to incorporate multicollinearity into their cue-criterion weights, instead of basing weights on cue-criterion relations only and later making imperfect adjustments to those weights or their judgments. This focus and

explicit attention on multicollinearity is expected to reduce attentional and attentional-by-processing errors.

Further, a construct structure may make cognitive processing easier by decomposing the task into two distinct types of relations that are estimated separately, reducing processing error in those estimates (Payne et al. 1992). Overall, total judgment error will be reduced and there will be smaller differences in total judgment error in tasks with two than five cues and with low than high multicollinearity. The complete form of the ordinal interaction of the number of cues, multicollinearity, and task structure on total error is shown in both graphs in Figure 2.

By decomposing the task with a construct structure, individuals may also be prompted to examine differences between any initial beliefs about relations among the cues and criterion and their beliefs after estimating the value of the organizational objective. This may make individuals more aware of inconsistencies in their thinking so that they can focus more clearly on developing and using a single judgment heuristic, thus improving judgment consistency (Ashton 1990; Goodwin and Wright 1994; Messier 1995).

In summary, with the indicator structure, increases in the number of cues and increases in their multicollinearity are expected to interact ordinally to increase total error and thus decrease judgment performance, but these negative ordinal interactive effects are expected to be reduced with a construct structure. The interactive effect of the number of cues, multicollinearity, and task structure on judgment performance is graphed in Figure 3 (note that Figure 3 is the complement of Figure 2, which has total error as the

dependent variable rather than judgment performance). The following general hypothesis states these effects.

General Hypothesis:

When making judgments of a criterion using a set of cues, there will be a three-way interaction between the number of measures, multicollinearity, and task structure on judgment performance, as follows:

(a) with either an indicator or a construct structure:
- judgment performance will be lower when there are five cues than when there are two cues (for indicator structure, 1>5 and 3>7; for construct structure, 2>6 and 4>8); and,
- when there are five cues as opposed to two cues, the difference in judgment performance between judgments made with cues with high multicollinearity and cues with low multicollinearity will be larger (for indicator structure, (5-7)>(1-3); for construct structure, (6-8)>(2-4)).

(b) comparing judgment performance across an indicator versus a construct structure:
- holding the number of cues and multicollinearity constant, judgment performance will be higher with a construct structure than with an indicator structure (2>1, 4>3, 6>5, 8>7);
- holding the number of cues constant, the difference in judgment performance that results when using cues with high multicollinearity as opposed to cues with low multicollinearity will be smaller with a construct structure than with an indicator structure ((1-3)>(2-4) and (5-7)>(6-8)); and,
- the ordinal interaction between the number of cues and their multicollinearity (described in the second bullet in (a) above) will be larger with an indicator structure than with a construct structure ((5-7)-(1-3) > (6-8)-(2-4)).

INSERT FIGURE 3

Dimensions of Judgment Performance

Accuracy of an individual's estimated cue-criterion weights measures how closely the OLS weights in an individual's policy-capturing model correspond to the OLS weights in the environmental model for the task (as defined earlier, just before the "Processing Error and Number of Cues" subsection). How accurately individuals

estimate the relations in an environment is a foundation of accurate judgment, which has a direct economic impact in organizations. If individuals do not use the incremental information in a given measure in their judgments, then the quality of their judgments will be lower and the resources invested to collect that information will be wasted. As discussed earlier, it is expected that with the indicator structure, increases in the number of cues and increases in their multicollinearity will ordinally interact to decrease accuracy in cue-criterion weights, but a construct structure is expected to reduce those negative ordinal interactive effects. Hypothesis 1 is a formal statement of these effects for accuracy of estimated cue-criterion weights, which are illustrated in Figure 3.

H1: The accuracy of estimated cue-criterion weights is a three-way interaction of the number of measures, multicollinearity, and task structure.

Judgment consistency measures how invariant an individual is in applying the OLS weights in his or her policy-capturing model. When an individual's judgments are the basis for resource allocation decisions for a program directed at changing the level of some organizational objective, variation in those judgments can lead to overspending on the program at some times and underspending at others. In addition, inconsistency in judgments may be construed as a signal that an individual is unsure of the true relations in the cues and criterion he or she is using and is trying to "play it safe" by using different cues for different judgments. This has two implications. First, depending on the extent of variation in judgments, the benefits of using some type of decision aid to improve judgment performance may exceed the cost. Second, if the individual continues to have difficulty estimating the relations in the environment and tends to habitually use different cues for different judgments, he or she may not detect when critical causal relations change.

As discussed earlier, it is expected that with an indicator structure, the number of cues and multicollinearity will interact ordinally to decrease judgment consistency, but a construct structure will reduce those negative ordinal interactive effects. Hypothesis 2 formally summarizes the expected effects for judgment consistency, which are illustrated in Figure 3.

H2: Judgment consistency is a three-way interaction of the number of measures, multicollinearity, and task structure.

CHAPTER 3: EXPERIMENTAL DESIGN

This chapter begins with a description of the participants in the experiment and a power analysis based on the number of participants. Subsections that follow include the independent variables, the experimental setting and the procedures followed in the administration of the experiment, and the dependent variables.

Participants and Power Analysis

The 101 participants in the experiment were 10 Ph.D. students, 69 first-year MBA students, and 22 upper-level undergraduate students who served as teaching assistants for two introductory accounting courses. Participants were paid performance-contingent compensation, as described in the "Procedures" section below.

A power analysis was conducted to determine the sample size needed to detect significant effects given the experimental design. Based on an estimated population effect size of R^2=0.15, for power of 90%, 112 participants were needed in total. While the actual number of participants of 101 is slightly lower than the required number, the average R^2 of the subsequent ANOVA's was 0.25. Thus, power appears to be satisfactory.[6]

Independent Variables

The experimental design is a 2x2x2 between-subjects factorial. The first independent variable is the number of measures of the product quality objective, either two or five. The second independent variable is the multicollinearity in the measures of product quality, either low or high. The difference in low and high multicollinearity is based on a series of *Z*-tests which compare the pairwise correlations between the product quality measures in the low and high multicollinearity data sets; the pairwise correlations

between measures in the low multicollinearity conditions are significantly lower ($p<.10$, except for two which are significantly lower at $p<.20$) than those in the high multicollinearity conditions (see Table 1 for statistics on the data used in the task, including correlation matrices). While the assessed difference in low versus high multicollinearity is based on these *Z*-tests, a visual inspection of collinearity diagnostics (Table 1, Panel D) also provides an indication that there are differences in multicollinearity in the low versus high conditions.

INSERT TABLE 1

The third independent variable is task structure, which requires participants to either make predictive judgments of measures of customer satisfaction (indicator structure), or to make judgments of both overall product quality and measures of customer satisfaction (construct structure). These two task structures are included in Appendix D, the "Envelope 2" subsection.

Experimental Setting

Participants are told that they are managers in an organization in which upper management believes that product quality affects customer satisfaction. The organization is implementing a new performance measurement system in which both product quality and customer satisfaction objectives have multiple measures; a diagram of the new performance measurement system is then provided (Figure 4). A complete set of experimental materials is provided in Appendix D.

INSERT FIGURE 4

Participants are told that management is interested in how the particular performance measures they chose to use in the new system help them learn the relation between product quality and customer satisfaction, so they will be asked to make predictive judgments using the new measures. They are next given information about the performance measures for product quality and customer satisfaction. To control for the possibility that participants' prior beliefs about relations between specific measures of product quality and customer satisfaction would affect their judgments (Miller 1971; Broniarczyk and Alba 1994; Luft and Shields 2001), all measures have generic labels (e.g., product quality measure #1, customer satisfaction measure #1, etc.). Further, to control for differences in judgments that might result if the performance measures had different metrics (Tversky et al. 1988; Slovic et al. 1990), all measures are scale-free and transformed to have the same means and standard deviations (p>.60).[7]

Participants then receive information about their organization. They are told that there are 40 plants in the organization that all make the same product and are built to the same design, so the production scale and technology is similar in all of them. In addition, the customers served by each plant are similar. Because of these similarities in product, production scale, technology, and customers, the effect of product quality on customer satisfaction is roughly the same across plants, but there are minor between-plant differences that could cause variation in the effects of product quality on customer satisfaction. Participants are also told that there are no shocks or seasonal variations in

the data that might cause variations in the relation between product quality and customer satisfaction.

After this introduction, participants in the indicator structure condition are given a table with past observations of the performance measures for product quality and customer satisfaction for 20 of the 40 plants (the "learning data set"; see Table 2 for an example), and are told to study the data until they believe they understand the relation between product quality and customer satisfaction. Participants in the construct structure condition are first given a table of past observations of only the product quality performance measures for 20 of the 40 plants, and are asked to estimate the level of product quality for each of the same 20 plants (participants are not asked to make judgments of the level of customer satisfaction to keep task requirements within the time available). They are then given the same table as participants in the indicator structure condition, which includes past observations of the performance measures for both product quality and customer satisfaction; the product quality data on this table are the same as that on the table used to make the estimates of the level of product quality. Like the participants in the indicator structure condition, they are told to study the data until they believe they understand the relation between product quality and customer satisfaction (see Appendix D, "Envelope 2" subsection for differences in these task structures).

INSERT TABLE 2

Much of the prior research in the policy-capturing paradigm is designed so that participants are given the opportunity to learn relations in data by providing them with case-by-case feedback; they are shown the actual value of the criterion for a given set of cues after they have made their prediction of the criterion but before they make their prediction of the criterion for the next set of cues (see Libby 1981, p. 29, Ashton 1982, p. 33, and Klayman 1988 for a description of this design). In contrast, in this experiment the opportunity to learn relations in the data is provided by giving participants all of the cue-criterion cases at once in tabular form, without the requirement that predictions be made first (see Hutchinson and Alba (1997) and Luft and Shields (2001) for examples of this design). With both the case-by-case and tabular designs, individuals have the same information and are provided with feedback about outcomes (i.e., the value of the criterion); the difference is whether those outcomes are provided on a case-by-case basis or simultaneously.

The tabular design is used in this experiment for four reasons. First, prior research finds that judgments of relations between cue and criterion variables are not different for individuals who make case-by-case predictions than for individuals who examine all cases at once (Well et al. 1988),[8] and one study suggests that individuals learn better by watching others do a task than by doing it themselves (Merlo and Schotter 2001). Second, if use of a tabular design does in fact inhibit learning, then it would simply lower the means of the dependent variables for each experimental condition, but there is no reason to expect that it would change the differences in means across experimental conditions. Therefore, conclusions about differences that arise from the effects of number of measures, multicollinearity, and task structure on judgment

performance would still be valid. Third, the tabular design is representative of the data available for common business judgment tasks (Hutchinson and Alba 1997), particularly those in which individuals learn from the experiences of others rather than from their own experiences (e.g., a corporate-level manager who learns about the performance of a division based on reports from each subunit in that division, and then makes predictions about divisional performance). In practice, it is frequently upper-level managers who are interested in the causal relations between measures of the operations of different departments within their organization (as opposed to departmental managers who typically do not have access to performance measures for other departments), so the tabular presentation of information is appropriate for examining judgment performance with respect to estimating and applying these relations. Fourth, prior research in accounting (Luft and Shields 2001) and psychology (Hutchinson and Alba 1997) has used this design for business judgment tasks.

Policy-capturing studies often require participants to make 100 to 200 case-by-case judgments which are then divided into blocks, with the first blocks considered the learning phase (e.g., Naylor and Schenck 1968; Armelius and Armelius 1974; Schmitt and Dudycha 1975). In contrast, in this experiment participants are given a table of 20 observations from which to learn the relations in the data, and then make 20 predictive judgments. A lower number of judgments is used in this experiment for three reasons. First, Brehmer (1987) suggests that learning in predictive judgment tasks takes place rapidly or not at all. Second, prior research used abstract judgment tasks with few supplemental questions included in the experiment, while this experiment includes extensive pre- and post-experimental questionnaires. Therefore, in the interest of keeping

the task length within the time available, 20 trials are used. Third, an increase in the number of trials would simply increase the means of the dependent variables for each experimental condition, but there is no reason to expect that it would change the differences in means across experimental conditions. Therefore, conclusions about the effects of number of measures, multicollinearity, and task structure on judgment performance would still be valid.

The learning data for each of the four experimental conditions are generated with a computer program that used experimenter-specified parameters as inputs. Care was taken to control that the realized parameters of the four learning data sets differed from each other with respect to the number of measures and multicollinearity only (realized parameters of each of the four learning data sets are in Table 1). Specifically:

- the adjusted-R^2 of the four regression models of customer satisfaction measure #1 on product quality are comparable to each other, ranging from .71 to .74 (Table 1, Panel A);
- the adjusted-R^2 of the four regression models of customer satisfaction measure # 2 on product quality are comparable to each other, ranging from .57 to .65 (Table 1, Panel A);
- the means and variances of the product quality and customer satisfaction measures do not differ across the four data sets (p>.60) (Table 1, Panel A);
- the bivariate correlations between each product quality and customer satisfaction measure and between the two customer satisfaction measures do not differ from each other across data sets (p>.20) (Table 1, Panel C);

- the bivariate correlations between pairs of product quality measures in the high and low multicollinearity conditions do differ from each other ($p<.10$, except for two comparisons which differed at $p<.20$) (Table 1, Panel C);
- for a given number of measures, the highest VIF, mean VIF, and condition index are higher for the high than the low multicollinearity data sets (Table 1, Panel D).

After studying the learning data without the aid of a calculator, participants are given a table with product quality measures for the 20 other plants in their organization (the "judgment data set"). For each of the potential levels of the product quality measures, the participants are asked to make predictive judgments of the customer satisfaction measures that they expected would result, given the levels of the product quality measures. The format of the table provided for the judgment task is identical to that in Table 2, except that the customer satisfaction columns are left blank.

To control for differences between the learning and judgment data sets that could affect judgment performance, the judgment data for each of the experimental conditions are obtained by applying a transformation to the product quality and customer satisfaction values from the learning data set so that the means and standard deviations are slightly different but other realized parameters remain the same (see Table 1, Panels A through D for statistics for the data sets).[9] After the transformation, the means and standard deviations of the measures across the learning and judgment data sets do not significantly differ from each other ($p>.60$), the correlation matrix and regression weights are the same, and all adjusted-R^2, correlation, and collinearity diagnostics comparisons described above are the same.

It is important to note that for the five measure-high multicollinearity data set, the OLS regression weight on product quality measure #4 is negative and significant ($p<.05$) in the environmental model with customer satisfaction measure #1 as the dependent variable, as is the OLS regression weight on product quality measure #3 ($p<.10$) in the environmental model with customer satisfaction measure #2 was the dependent variable (see Table 1, Panel B). As the number of measures and their multicollinearity increase, it is likely that OLS regression weights for some product quality measures will be negative although their bivariate correlations with the customer satisfaction measures are positive.[10] Therefore, although the negative OLS regression weights differentiate the five measure-high multicollinearity data set from the others, this is likely to be representative of data drawn from the natural ecology. A more thorough discussion of the effects of multicollinearity on OLS regression weights is presented in Appendix B.

Procedures

Participants reported to a classroom and were randomly assigned to one of the experimental conditions. The experiment was administered with paper and pencil materials. Four envelopes of materials were at each participants' seat, and they were able to self-pace their way through them (a complete set of experimental materials is provided in Appendix D). Participants took an average of 45 minutes to complete the experiment.

Upon being seated, the participants were informed of the compensation system. They were paid contingent on the accuracy of their predictive judgments of customer satisfaction. Pay ranged from $10 to $20 per person. A quadratic loss function was used to compute judgment accuracy relative to the best possible judgments that could be made (i.e., predictions using the environmental model, computed by applying the OLS weights

from a regression model of the learning data set to the values of the product quality measures in the judgment data set). For each participant, an error measure was computed and summed across all 20 judgments, using the formula (your judgment - best possible judgment)^2. Cash payment was linearly inversely related to the magnitude of the error measure.

After learning the compensation system, participants began to work on the materials in first envelope, in which they were asked questions intended to capture their prior beliefs about the relationship between product quality and customer satisfaction and their beliefs about the overall importance of product quality and customer satisfaction. After they returned these materials to the first envelope, they opened the second envelope and reviewed the learning materials, and then made judgments of the product quality construct (for the construct structure condition only) that were previously described. Once they completed the learning materials, they moved on to the third envelope which contained the judgment task materials, but they were allowed to keep the learning materials accessible. After completing the judgment task, they were asked a series of questions on the just-completed task (confidence in the judgments they just made; difficulty, complexity and familiarity of the task; self-assessed weights placed on the product quality measures during the judgment task), and they returned both the learning and task materials to an envelope. The fourth envelope contained the post-experiment questionnaire, which asked questions about participants' cognitive judgment heuristic, assumptions they made when using the performance measurement system, familiarity with performance measurement systems, statistical knowledge, and demographic information.

Dependent Variables

Since this dissertation investigates judgment performance with respect to subjective estimation and application of cue-criterion weights, not the accuracy of the resulting predictions themselves, lens-model correlational measures that are computed using predictions of the criterion are not used as dependent variables (e.g., Libby 1981; Ashton 1982; Luft and Shields 2001). Further, were the lens-model correlational measures used, their values would be inflated for the high multicollinearity conditions.

As noted previously, prior literature shows that the correlation between predictions produced by applying different weights to the same set of cues is an increasing function of the number of cues and their multicollinearity (Libby 1981, p. 42; Ashton 1982, p. 37). In other words, if two individuals apply different weights to the same cues, and those cues are highly multicollinear, then the predictions of the individuals will be highly correlated despite the fact that they use different cue weights to produce those judgments. For example, if one individual places a large weight on the first available cue and small weights on the remaining cues, while the other individual does the opposite, then the predictions of the two individuals will be highly correlated because of the multicollinearity in the cues, even though they use different cue weights in their policy-capturing model. This dissertation uses dependent variables based on OLS estimates of cue-criterion weights from individuals' policy-capturing models to try to capture differences in how well individuals estimate weights; these dependent variables are affected less by the problems associated with multicollinearity than the lens-model dependent variables based on individuals' predictions of the criterion (as is discussed further below and in Appendix C).

As a basis for computing the dependent variables, a participant's policy-capturing model is estimated by regressing his or her predictions of customer satisfaction on the product quality measures used to make the predictions (i.e., from the judgment data set).[11] The environmental model of the task is determined by regressing the customer satisfaction measures on the product quality measures from the learning data set. The dependent variables are then based on computations using the OLS regression weights from these models.

For H1, the accuracy of an individual's estimated cue-criterion weights is computed as the mean absolute difference between the OLS regression weights from the participant's policy-capturing model and the OLS regression weights from the environmental model of the task. This computed value is subtracted from one so that the higher the value of the dependent variable, the closer an individual's estimated weights are to the OLS weights in the environmental model (i.e., a value of one for this dependent variable means that the weights in an individual's policy-capturing model equal those in the environmental model). The mean absolute differences are averaged to make the dependent variable comparable across the two-measure and five-measure conditions. The mean absolute difference between the OLS weights in the policy-capturing and environmental models, rather than the mean signed difference or the mean relative difference (i.e., the difference in the policy-capturing and environmental OLS weights as a percentage of the environmental OLS weights), is used to compute this dependent variable. Use of the mean signed difference in OLS weights would allow errors in cue-weights to offset each other, but any less-than-accurate use of incremental information decreases the accuracy of judgments, and if those judgments are used to make resource

allocation decisions there is a real economic impact to the organization. Further, the resources used to collect performance measure data that individuals do not use will be wasted. Use of the mean relative difference in OLS weights would lead to problems with comparability since the OLS weights in the environmental model differ across experimental conditions (i.e., the mean relative difference will be different for an error of the same size across different conditions).

For H2, the degree to which an individual consistently applied the cue-criterion weights in his or her policy-capturing model is computed as the mean absolute difference between the OLS regression weights from a policy-capturing model of the participant's first seven judgments of customer satisfaction and a policy-capturing model of the participant's last seven judgments. This computed value is subtracted from one so that the higher the value of the dependent variable, the more closely an individual's estimated weights for the first seven judgments are to those used for the last seven judgments (i.e., a value of one for this dependent variable means that the estimated cue-criterion weights the individual used in the first seven and the last seven judgments are equal). The first seven and last seven judgments are used to capture the difference in subjective weights for the first third and the last third of the required judgments (the judgments are broken down into thirds because at least seven judgments are required to estimated policy-capturing models for participants in the five-measure condition). The differences in the weights are averaged to make the dependent variable comparable across the two-measure and five-measure conditions.

The variance of estimated OLS regression weights is higher when there is high multicollinearity in independent variables. Consequently, it is possible that the

dependent variables for the high multicollinearity experimental conditions are measured with more error. If this difference in measurement error across the high and low multicollinearity conditions is significant, then it will result in a violation of the homogeneity of variance assumption of regression and ANOVA, since the variance of error terms will be higher in the high multicollinearity experimental conditions than in the low multicollinearity conditions. Based on Levene's test, the variances of the error terms do not significantly differ for the measure of accuracy of estimated cue-criterion weights ($p>.05$), but differ significantly for the measure of judgment consistency ($p<.05$). A $\log_{10}$ transformation of the judgment consistency measure eliminates this violation ($p<.10$), but results of an ANOVA using the transformed variable are not qualitatively different than results using the raw variable.[12] Therefore, hypothesis tests are conducted using the raw dependent variables.[13] Further discussion and analysis of the effects of multicollinearity on the dependent variables is in Appendix C.

CHAPTER 4: RESULTS OF EXPERIMENT

This chapter provides evidence on whether participants used the heuristics assumed, and tests of whether differences in participants across experimental conditions affected the results. Results of hypothesis tests and supplemental analyses follow.[14]

Analysis of Heuristics Used

In the post-experimental questionnaire, participants were asked to respond to a series of questions designed to determine any heuristic(s) they used when making their customer satisfaction judgments. Responses to those questions were then classified into several categories based on the heuristics represented. Of the 101 participants:

- 67 participants (66.3%) indicated that they used either the difference heuristic, the equal-weight heuristic, or both (because of the design of the post-experiment questions, it is difficult to determine with confidence how many of the 67 participants fell into each of these three categories);
- 20 participants (19.8%) indicated that they relied exclusively on an exemplar heuristic (see Appendix A for a description of this heuristic);
- two participants (2.0%) indicated that they computed the mean of customer satisfaction from the learning data and made adjustments to that value for each judgment (this is analogous to a chunk-based heuristic described in Hutchinson and Alba (1997), in which individuals combine observations in some manner and determine general trends before making judgments; none of the participants in the small-scale empirical investigation detailed in Appendix A used this heuristic);
- two participants (2.0%) used the same value for each of their judgments of customer satisfaction; and,

- it was unclear from the responses what heuristic the remaining ten participants (9.9%) used.

The same post-experiment questions were examined to determine whether the 67 participants who used the difference and/or equal-weight heuristics switched between them in the course of making their judgments because of use of an exemplar-based heuristic. Of those 67 participants, 62 (61.4% of the total sample) switched. These results were consistent with the small-scale empirical investigation of potential heuristics used to complete this task reported in Appendix A.

Tests of Randomization and Sensitivity of Results

To test whether differences across participants may have driven any results, measures of the following were included as both as the dependent variables in 2x2x2 ANOVA's and as covariates in separate 2x2x2 ANCOVA's for each dependent variable:

- participants' prior beliefs about the relation between product quality and customer satisfaction, the importance of product quality personally and economically, and beliefs about and experience with strategic performance measurement systems;
- measures of participants' knowledge of accounting, statistics, SEM, math, finance, supply chain, quality management, and operations management; and,
- responses to questions about the complexity of the experimental performance measurement system, and difficulty and familiarity of the task itself.

The results of the ANOVA's indicate that characteristics of the participants (e.g., prior beliefs, knowledge, or experiences) did not significantly differ ($p>.05$) across the experimental conditions. The results of the ANCOVA's for both the accuracy and consistency dependent variables and the resulting patterns of adjusted means were not

qualitatively different than those based on the ANOVA's. Therefore, random assignment to the experimental conditions appears to have been successful, and the tests that follow exclude covariates.

Tests of Hypotheses and Supplemental Analyses

To test each hypothesis, a series of 13 planned contrasts was performed based on the predicted pattern of means shown in Figure 3. A Bonferroni adjustment was used to control family-wise error at $p<.05$, so the significance level for each individual contrast within a given hypothesis was $p<.004$. Results of ANOVA's are also presented in the tables for each hypothesis test.[15]

Test of H1: Accuracy of Estimated Cue-Criterion Weights

The accuracy of estimated cue-criterion weights measured the degree to which the OLS cue-criterion weights in an individual's policy-capturing model differed from those in the environmental model for the task. Descriptive statistics for this dependent measure are in Table 3, Panel A.

INSERT TABLE 3

Contrast tests were based on the predicted pattern of cell means in Figure 3; the results are graphed in Figure 5 and presented in Table 4, Panel A. Of the 13 planned contrasts, three were in the direction predicted and significant at $p<.004$ (Table 4, Panel A, test numbers 2, 5, and 6). Comparing the indicator to the construct conditions, no contrasts were significant ($p>.15$; Table 4, Panel A, test numbers 7 through 13), but there

were significant contrasts within the indicator structure and within the construct structure conditions themselves.

INSERT FIGURE 5

INSERT TABLE 4

Within the indicator structure condition (Figure 5), the overall pattern of means was close to that predicted, although accuracy was lower than expected for condition 1 which resulted in no significant difference ($p>.40$) in the low-multicollinearity conditions (Table 4, Panel A, test number 1). The means for the high-multicollinearity conditions were significantly different at $p<.004$ (Table 4, Panel A, test number 2). In addition, the number-of-measures-by-multicollinearity contrast was significant ($p=.008$) at slightly higher than the Bonferroni-adjusted level (Table 4, Panel A, test number 3).

Within the construct structure condition (Figure 5), the overall pattern of means for the high multicollinearity condition was as predicted, but not for the low multicollinearity condition. The means for the high multicollinearity conditions were significantly different at $p<.004$ (Table 4, Panel A, test number 5), and the number-of-measures-by-multicollinearity interaction was significant at $p<.004$ (Table 4, Panel A, test number 6). Contrary to predictions, however, mean accuracy for the five-measure-low-multicollinearity condition was significantly higher than that for the two-measure-low-multicollinearity condition at $p<.004$ (Table 4, Panel A, test number 4).

Discussion of Results for H1: Accuracy of Estimated Cue-Criterion Weights

Overall, the results of the planned contrasts provided partial support for H1. The

number-of-measures-by-multicollinearity interactions within the indicator structure and within the construct structure conditions (Table 4, Panel A, test numbers 3 and 6, respectively) were significant at $p<.01$. An increase in the number of measures resulted in lower accuracy only when multicollinearity in the measures was high. However, the three-way interaction was not significant at $p>.20$ (Table 4, Panel A, test number 13), indicating that the effects of the number of measures and multicollinearity on judgment performance did not differ across different task structures.

The results of the planned contrasts were consistent with the results of an ANOVA with accuracy as the dependent variable and the number of measures, multicollinearity, and task structure as the independent variables (Table 4, Panel B). The ANOVA showed a significant number-of-measures-by-multicollinearity interaction ($F=17.23$, $p=.00$), but no significant main or interactive effects for task structure ($p>.20$), despite the higher-than-predicted accuracy for the five-measure-low-multicollinearity-construct-structure condition.

A potential explanation for the ineffectiveness of the construct task structure to lead to higher accuracy was that it did not focus attention on multicollinearity as assumed it would. To check for this possibility, a post-task question asked participants to rate their extent of agreement with the statement, "I thought that some or all of the product quality measures I used to make my estimates were highly correlated with each other," using a Likert scale of 1="strongly disagree" to 10="strongly agree". It was expected that the mean response to this question would be higher for the high-multicollinearity-construct-structure condition than for the other conditions. When multicollinearity was low, there should have been no differences in mean responses between the indicator and

construct structure conditions, since calling attention to multicollinearity should have made no difference to assessments of its level when the level was low. When multicollinearity was high, however, it was expected that the mean response to this question would have been higher for the construct than for the indicator structure condition. The mean response of 6.52 in the high-multicollinearity-construct-structure condition was higher than the mean response of 5.75 for the other three multicollinearity-by-task-structure conditions ($t>1.68$, $p<.05$, one-tailed). This was consistent with the assumption that the construct structure did in fact focus attention on multicollinearity.

A further post-task question was designed to disentangle whether judgment performance was not affected by the construct structure because participants did not understand that multicollinearity affected estimates of the product quality-customer satisfaction relations, or because they understood that multicollinearity was important to estimates of those relations but did not know how to incorporate it into their judgments. Using a Likert scale of 1="strongly disagree" to 10="strongly agree", participants answered the question, "Although the relationships **between** the product quality measures themselves have an impact on the relationships between the product quality and the customer satisfaction measures, I do not know how to incorporate this into my estimates." It was expected that the mean response to this question would have been higher for the high-multicollinearity-construct-structure condition than for the other conditions since participants in that condition should have been prompted to focus on multicollinearity and thus should have more consciously attempted to incorporate it into their judgments. The mean response for the high-multicollinearity-construct-structure condition of 7.30 was significantly higher than the mean of 5.27 for the other three

multicollinearity-by-task-structure conditions (t>2.60, p<.01, one-tailed). This provided further support for the assumption that the construct structure did focus attention on multicollinearity, but it appears that even when prompted to focus on multicollinearity, individuals did not know how to incorporate it into their judgments. This was also supported by the fact that in the small-scale empirical investigation that was conducted (Appendix A), one participant who did focus on multicollinearity, albeit without being prompted to do so, said he did not know how to incorporate it into his judgments.

Test of H2: Judgment Consistency

Judgment consistency measured the degree to which an individual consistently applied the weights in his or her policy-capturing model to the measures in the task. Descriptive statistics for this dependent measure are in Table 3, Panel B. Contrast tests were based on the predicted pattern of cell means in Figure 3; the results are graphed in Figure 6 and presented in Table 5, Panel A. Of the 13 planned contrasts, one was significant at p<.004 (Table 5, Panel A, test number 5).

Comparing the indicator to the construct conditions, there were no significant contrasts at the Bonferroni-adjusted level (Table 5, Panel A, test numbers 7 through 13). Within the indicator structure and within the construct structure conditions themselves, two contrasts were significant (one at the Bonferroni-adjusted level, one at p<.05).

INSERT FIGURE 6

INSERT TABLE 5

Within the indicator structure condition, the overall pattern of means was consistent with the prediction (Figure 6), although the number-of-measures-by-multicollinearity interaction was not significant at $p>.20$ (Table 5, Panel A, test number 3). At a significance level of $p<.05$, however, the contrast comparing the high-multicollinearity conditions (Table 5, Panel A, test number 2) indicated that the mean for condition 7 was significant lower than that for condition 3.

Within the construct structure condition, the overall pattern of means was as predicted (Figure 6), although the number-of-measures-by-multicollinearity interaction was not significant at $p>.05$ (Table 5, Panel A, test number 6). The contrast comparing the high-multicollinearity conditions (Table 5, Panel A, test number 5) indicated that the mean for condition 8 was significantly lower ($p<.004$) than that for condition 4.

Discussion of Results for H2: Judgment Consistency

Overall, these results of the planned contrast tests provided little support for H2. The number-of-measures-by-multicollinearity interaction within the indicator structure condition (Table 5, Panel A, test number 3) was not significant, and was only marginally significant ($p<.08$) within the construct structure condition (Table 5, Panel A, test number 6). However, there were significant differences in judgment consistency between the high multicollinearity conditions within both the indicator and the construct structure conditions.

The results of an ANOVA (Table 5, Panel B) indicated that there were significant main effects for the number of measures ($F=14.18$, $p=.00$) and multicollinearity ($F=22.81$, $p=.00$) on judgment consistency, but the number-of-measures-by-multicollinearity interaction was not significant ($p>.10$). The significant main effects in

the ANOVA, coupled with the visual inspection of the obtained pattern of means in Figure 6, revealed that both the number of measures and multicollinearity reduced judgment consistency, but did not do so interactively. This effect visually appeared to be more striking in the construct than in the indicator structure condition (Figure 6), but it was not so great to lead to significant main or interactive effects of task structure on judgment consistency.

With respect to task structure, the contrast test for the three-way interaction was not significant (Table 5, Panel A, test number 13), and there were no main or interactive effects of task structure ($p>.05$) in the results of the ANOVA (Table 5, Panel B). These results indicated that differences in task structure had no effect on judgment consistency. As noted in the discussion of the results for H1 above, the lack of effects of task structure appeared to be due to the fact that individuals did not know how to incorporate multicollinearity into their subjective judgments even when they were prompted to do so.

CHAPTER 5: DISCUSSION AND CONCLUSION

This dissertation empirically investigated how the use of multiple performance measures affects individuals' judgment performance. Specifically, it provided theory-based experimental evidence on how the number of performance measures used to measure a particular organizational objective and the multicollinearity in those measures affect individual judgment performance in a prediction task. Further, it investigated how a change in the structure of the task can influence judgment performance. Measures of judgment performance were intended to capture how accurately individuals estimated the relations between and among performance measures, and how consistently they applied the relations they estimated to make predictive judgments. This chapter synthesizes the results of the experiment and their relation to the hypotheses and to prior research, discusses the contributions and limitations of this dissertation, and provides possible directions for further research.

Synthesis of Results

Results of this experiment were only partially supportive of the 13 predicted judgment performance differences for each of the two hypotheses. Of those 13 predictions for each hypothesis, six related to the effects of the number of performance measures and their multicollinearity on judgment performance, and seven related to the effects of task structure on judgment performance.

Results of contrast tests for the former six predictions indicated that judgment performance was a function of the number of performance measures used to measure an organizational objective and their multicollinearity. With respect to the accuracy with which individuals estimated relations between and among performance measures, there

was an interactive effect of the number of measures and multicollinearity. An increase in the number of measures did not result in significantly less accurate estimates of the relations if the multicollinearity in those measures was low, but did so when multicollinearity was high. With respect to the consistency with which individuals applied their estimated relations to a prediction task, both increases in the number of performance measures and increases in their multicollinearity led to lower consistency, but the effects were not interactive as was predicted.

Results of prior research that focus on the number of measures that cause individuals to reach the point of information overload suggest that judgment performance does not begin to decrease until there are five or more quantitative measures in an information set (Tuttle and Burton 1999). While the information load research generally investigated judgment performance across sets of information with different predictive ability, or changes in judgment performance as measures were incrementally added to a set, it did not address the effects of multicollinearity in the measures. The results of this dissertation suggested that decreases in judgment performance may occur with five measures if those measures are multicollinear.

Consistent with the results of this dissertation, prior research using a variety of judgment tasks and judgment performance measures indicate that individuals did not typically incorporate multicollinearity into their judgments, resulting in lower judgment performance when multicollinearity was high (Armelius and Armelius 1974; Brehmer 1974b; Lindell and Stewart 1974; Schmitt and Dudycha 1975; Libby 1981; Schum and Martin 1982; Klayman 1988; Maines 1990, 1996). Two of those studies used tasks similar to that in this dissertation, and found that judgment performance with respect to

the accuracy of estimated cue-criterion weights (as measured using the lens-model matching index) was lower when there were only two multicollinear cues (Lindell and Stewart 1974; Schmitt and Dudycha 1975). However, the results of this dissertation suggested that accuracy was not lower with two multicollinear cues, but was with five.

There are three possible reasons for the difference in results for the two-cue task in this dissertation and for prior research using similar two-cue tasks. First, the dependent measures of accuracy differ; this study used a measure based on OLS weights, while the prior studies used the lens-model matching index. Second, the additional contextual features of this task, as opposed to the abstract nature of the tasks in the prior studies, may have led to higher judgment performance by engaging participants in the task to a greater extent (Libby 1981, p. 30). Third, the structure of the performance measurement system in the experimental materials may have given participants a cue that the measures were multicollinear, and the participants were able to effectively incorporate the multicollinearity into their subjective judgments at the two-cue level but not at the five-cue level (see the "Limitations" section below for a further discussion of this issue). The third possibility is consistent with results in Lipe and Salterio (2002), in which individuals' judgments changed when performance measures were categorized by organizational objective.

With respect to the latter seven predictions about the effects of task structure, results of contrast tests showed that, overall, task structure had no effect on judgment performance. Use of a task structure designed to focus individuals' attention on multicollinearity and decompose cognitive processing requirements did not result in higher accuracy or consistency regardless of the number of measures and their

multicollinearity. Analyses of post-task questions suggested that when individuals were prompted to focus on multicollinearity they were aware that it should affect their estimates of relations between and among performance measures, but they did not know how to incorporate it into their judgments. Jiambalvo and Waller (1984) found that decomposition of an audit task did not lead individuals to make different judgments than they did when the task was not decomposed, but they were unable to determine if the result was due to a failure of the decomposition to direct attention to the critical parts of the task or a failure of individuals to process information even when the decomposition focused attention on it. Results of this study suggest that Jiambalvo and Waller's (1984) result could be due to the latter.

Unlike the results in this experiment and in Jiambalvo and Waller (1984), Schum and Martin (1982) found that decomposition of a task did help individuals process multicollinearity. While the qualitative information in their task could be the reason for the differences in results, it is also possible that decomposition of a task improves judgment performance only under certain conditions. However, because of the limited research that examines decomposition for quantitative predictive judgment tasks and that compares performance on decomposed and non-decomposed tasks, those conditions are not immediately evident.

Limitations

This dissertation has six limitations. Four of these arise from choices made in the design of the experiment. The first is due to statistical characteristics of the data used in the task; the second and third are characteristics of the performance measures used in the task which were necessary to isolate the effects of the independent variables on judgment

performance; and the fourth is due to a characteristic of the context for the task. The fifth limitation is due to a manipulation that did not address all potential cognitive processing difficulties, and the sixth limitation is due to an inability to explain an unexpected result.

First, in the high multicollinearity conditions, the level of multicollinearity in the five-measure conditions was higher than that in the two-measure conditions (as measured by the collinearity diagnostics in Table 1, Panel D). In other words, while generating the data used in the task, it was impossible to construct data sets in which the multicollinearity diagnostics for the five-measure data set were the same as those for the two-measure data set without making some of the five measures uncorrelated with others in the set. Therefore, it is not clear whether the effect of high multicollinearity on judgment performance were due to an interaction with the number of measures, or whether it was simply due to the fact that multicollinearity increased as the number of measures increased. While constructing two- and five-measure data sets that had comparable levels of multicollinearity was virtually impossible, it is also likely that the data used in the experiment is representative of the statistical characteristics of performance measures organizations use in practice, particularly if they use performance measurement systems like the balanced scorecard (Kaplan and Norton 1992, 1993, 1996a-c, 2000, 2001).

Second, significant effort was made to include as many contextual features as possible in the task. However, it was necessary to use generic labels and the same scale for the performance measures since prior research indicates these factors can influence judgment performance (Miller 1971; Tversky et al. 1988; Slovic et al. 1990; Broniarczyk and Alba 1994; Luft and Shields 2001). In organizations, the performance measures

individuals use have labels and are of different scales, so the results of this experiment must be applied cautiously when predicting or explaining judgment performance based on performance measures with such labels and scales. However, to the extent that an organization is interested in judgment performance when individuals are using new performance measures with which they have no familiarity, or are analyzing potential effects of spending on new programs, the results of this experiment can be relevant.

Third, as mentioned in Chapter 3 ("Experimental Design" subsection), this experiment used a task in which individuals estimated relations in the data using all available observations at one time and then made predictive judgments, as opposed to making judgments on a case-by-case basis and estimating relations as they proceeded through the cases. The results of this study, therefore, should not be used to predict or explain judgment performance for the latter type of task (e.g., when a divisional manager reviews a report for a given period, makes predictive judgments for a subsequent period, receives actual outcomes for the subsequent period, and repeats the process).

Fourth, it is possible that there were no significant effects of task structure on judgment performance because of the context in which the judgment task was set. Specifically, individuals were told that the cues were proxies for product quality and the criteria were proxies for customer satisfaction, and were given a diagram of the performance measurement system showing these relations. Such a performance measurement system is representative of those used in practice (Kaplan and Norton 1992, 1993, 1996a-c, 2000, 2001). However, it is quite possible that this context alone prompted individuals to think about multicollinearity and adjust for it in their judgments, and thus the task structure manipulation may have been too weak to induce differences in

judgment performance. While the post-experiment questions discussed in Chapter 4 ("Discussion of Results for H1: Accuracy of Estimated Cue-Criterion Weights" subsection) suggest that this was not the case, it is quite possible that in another context the task structure manipulation would have had more of an impact on judgment performance.

Fifth, with respect to task structure, the results of this dissertation suggest that individuals can recognize multicollinearity when prompted to do so, but have difficulty incorporating it into their judgments. The task structure manipulation in this dissertation only included a component to help individuals recognize multicollinearity; it did not include a component designed to help individuals process it. Therefore, the question of what might help individuals process multicollinearity is left unanswered.

Sixth, judgment accuracy was much lower than expected for the two-measure-high-multicollinearity conditions when individuals made predictions for customer satisfaction measure #2, and it was not clear from the data what was driving this result (see Endnote 15). It is possible that the interactive effect of the number of performance measures and multicollinearity was influenced by some other variable that was not measured in this study, but it is not evident what that might be.

Contributions

This dissertation makes three contributions to existing literature in accounting and psychology. First, while performance measurement systems that map key organizational objectives in a cause-and-effect chain are increasingly popular in practice, there is limited research on the effects of the design of these performance measurement systems on individual judgments, and none of the existing research focuses on whether the number of

measures and multicollinearity interactively affect those judgments, holding the predictive ability of the set of measures constant, and how changes in task structure may improve such judgments. This dissertation answers a call by Sprinkle (2002) for further research in a managerial accounting setting that examines how the use of multiple performance measures affects individuals' ability to make organizationally-desirable decisions, and proposes a task structure designed to help individuals more effectively make such decisions.

Related to this contribution, if cause-and-effect performance measurement systems are designed so that different links in the chain have approximately the same predictive ability, then examining differences in judgment performance holding predictive ability constant provides insight into how judgments might differ at these different links because of differences in the number of measures and their multicollinearity. Prior research in information load investigated judgment performance across sets of information with different predictive ability, or the change in judgment performance as measures were incrementally added to a set (Casey 1980; Shields 1980, 1983; Iselin 1988; Chewning and Harrell 1990; Tuttle and Burton 1999). These studies often tried to determine the point at which individuals reached information overload. In contrast, this dissertation examines whether judgment performance can decrease even before information overload is reached, which is relevant to both designers and users of multiple performance measure systems.

Second, much of the prior research that examines individual predictive judgments with multicollinear data was conducted with large data sets (often with 200 or more observations) and more abstract tasks, and used dependent variables that were more

prone to measurement error because of multicollinearity. With respect to the large data sets and more abstract design of the task, Libby (1981, p. 30) states that such features can understate judgment performance because they omit important contextual details, so the task in this dissertation included as much contextual information as possible. For example, some prior studies used lines of differing lengths as the cues and criterion (e.g., Armelius and Armelius 1974, 1975), while others used two-digit numbers but did not add any other contextual information to the task (e.g., Naylor and Schenck 1968; Schmitt and Dudycha 1975). Further, the statistical properties of the data used in some of the prior research is likely to be less representative of actual data used in organizations than the data used here. Specifically, the correlation between some cues and the criterion was statistically near zero, although the cues were correlated with each other (Armelius and Armelius 1974, 1975). It is unlikely that organizations would choose or use performance measures for an organizational objective that were not causally linked to performance measures for another organizational objective to at least some extent. Finally, individuals in organizations rarely have data sets as large as those used in prior literature, and in fact may have ten or fewer observations with which to learn relations and make judgments. Therefore, judgment performance in an accounting context is likely to be quite different than that in prior research.

With respect to the dependent variables, prior literature used lens-model correlational measures of judgment performance were inflated by increases in the number of measures and their multicollinearity (Libby 1981, p. 42; Ashton 1982, p. 37). This dissertation used dependent variables that were less prone to these measurement

problems, and thus provided an alternate way to examine judgment performance in the presence of multicollinear data.

Third, cause-and-effect performance measurement systems strongly resemble structural equation models (SEM's), and no prior research was found that examined whether decomposing a judgment task into parts that resemble those of structural equations models resulted in different judgments than a non-decomposed task. Further, studies of whether task decomposition can improve performance in predictive judgment tasks are very limited but are relevant to many types of business tasks, particularly budgeting and forecasting (Goodwin and Wright 1993, 1994).

Implications for Practice

There are two implications of the results of this experiment for organizations in which multiple measures are used to measure organizational objectives. First, the finding that using more measures was not as detrimental to judgment accuracy in estimates of relations between and among performance measures when there was low multicollinearity as it was when there was high multicollinearity is important, given the prevalence with which the use of multiple measures is recommended in the literature. While concern has been expressed that the use of more performance measures can lead to lower judgment performance (Ittner and Larcker 1998), results of this experiment indicate this will not necessarily be the case if the organization is measuring relatively independent dimensions of a difficult-to-quantify organizational objective. Much of the literature, however, suggests that organizations use multiple measures to reduce noise in the measurement of difficult-to-quantify organizational objectives, which is precisely the measurement choice that can lead to high multicollinearity (Kaplan and Norton 1992,

1993, 1996a-c, 2000, 2001; Balkcom, Ittner and Larcker 1997; Lambert 1998; Sjoblom 1998; Stivers et al. 1998; Kaplan and Tempest 1999; Hertenstein and Platt 2000). The results of this study indicate that if organizations use multiple measures, particularly to reduce noise in measurement, individuals' judgment accuracy and consistency may be affected, and the use of some type of decision aid to help individuals process multicollinearity may be warranted.

Second, focusing attention on multicollinearity through the use of a different task structure did not affect judgment performance. It appeared that even if individuals' attention was focused on multicollinearity, they did not know how to incorporate it into their judgments. It is possible that some type of decision aid or task properties feedback might help individuals process multicollinearity, but what that might be other than use of a statistical model is not obvious. Further, given the reluctance individuals have to use statistical models to aid judgments (as discussed in Chapter 2, "Subjective Judgments in Organizations" subsection) it is not clear whether they would rely on the output from these models when making judgments. This, too, is an important finding given the trend to use multiple measures in performance measurement systems, and indicates that more research is needed to determine ways to help individuals both detect and process multicollinearity. Traditional methods of accounting and reporting may need to be redesigned to allow individuals to more effectively learn critical relations in organizational data if strategic performance measurement systems continue in popularity.

Possible Directions for Further Research

As noted about, this dissertation used dependent variables that were less prone to interpretation problems when there was high multicollinearity than were the lens-model

dependent variables that have been used in prior research. The dependent variables used here captured judgment performance based on individuals' estimates of the relations between variables (i.e., the OLS weights from their policy-capturing models), while those in lens-model research captured judgment performance based on individuals' predicted outcomes. Ashton (1981, p. 23) cautions against comparing results of studies which use different measures of judgment performance such as these. Therefore, one avenue for further research would be to reconcile differences in judgment performance based on dependent variables of the accuracy of predicted outcomes (e.g., prediction error), lens-model measures of accuracy and consistency, and the accuracy and consistency with which individuals estimated and applied the weights they used to predict those outcomes, all based on judgments made with the same data. This would present a clearer picture of different costs and benefits of using multiple performance measures that may or may not be highly correlated.

Little prior research has examined how individuals estimate relations between and among performance measures over an extended period of time, and whether the number of measures and their multicollinearity affect such judgments. Specifically, is there a point in time at which individuals begin to use only a subset of performance measures to make judgments, and is this point different if there are more measures in the set or if the measures are multicollinear? Will they switch between different subsets of measures across periods? Does the multicollinearity in the measures affect whether or not individuals will use a subset of the measures (i.e., if they detect multicollinearity, will they be more likely to use only a subset of measures)? If they use only a subset of measures to make judgments, then how well can they detect changes in the relations in

the data? If the same performance measures are used across multiple periods, then how often do individuals actually reestimate relations between measures to see if they have changed?

As noted previously, the results of this dissertation suggested that when attention was directed at multicollinearity, individuals knew that it was important to their estimates of relations between performance measures but they did not know how to incorporate it into their judgments. An interesting avenue for future research would be to examine whether other types of task decomposition that also call attention to multicollinearity, or whether task decomposition coupled with task properties feedback or task properties feedback alone, might lead to higher judgment performance.

Related to both of the latter suggestions for further research, Bonner's (1994) model of task complexity suggests that judgment performance will be higher if individuals simply reduce the number of measures to process by disregarding correlated measures, even though prior evidence indicates that individuals have difficulty disregarding information presented to them (e.g., Nisbett et al. 1981; Bloomfield et al. 1998a, b). Future research should investigate whether individuals disregard correlated measures if they are explicitly told to do so (which could be a type of task properties feedback and an attention-directing device), and whether doing so improves judgment performance. Further, it would be interesting to examine whether, over time, individuals would consistently disregard the same measures or if they would vary the measures they disregard to "hedge their bets", and the implications this has if relations in the performance measures changed.

Further, as noted earlier, there is limited prior research that examines the process individuals use to make predictive judgments like those examined in this dissertation, despite the fact that this is a common business judgment task. Research that attempts to predict and explain the process individuals use in predictive judgment tasks and how different factors affect that process (similar to the small-scale empirical investigation detailed in Appendix A) would provide insights into mechanisms that could be used to improve judgment performance.

TABLE 1

PARAMETERS FOR EXPERIMENTAL DATA SETS

PANEL A – Adjusted-R^2s of Customer Satisfaction-Product Quality Regressions and Means and Standard Deviations of Product Quality and Customer Satisfaction Measures

Data Set		Adjusted-R^2 of Regression Model, Customer Satisfaction On Product Quality		Learning Data Set -- Mean (std. dev.) of All Product Quality and Customer Satisfaction Measures [1]	Judgment Data Set -- Mean (std. dev.) of All Product Quality and Customer Satisfaction Measures [1]
No. of Measures	Multi-collinearity	Cust. Sat. Measure #1	Cust. Sat. Measure #2		
two	low	.72	.63	64.00 (6.40)	65.00 (6.50)
two	high	.71	.57	64.00 (6.40)	65.00 (6.50)
five	low	.74	.65	64.00 (6.40)	65.00 (6.50)
five	high	.74	.60	64.00 (6.40)	65.00 (6.50)

PANEL B - Regression Weights for Environmental Models of Customer Satisfaction on Product Quality [2]

Data Set			Regression Weight (p-value)				
No. of Measures	Multi-collinearity	Dependent Variable	Product Quality Measure #1	Product Quality Measure #2	Product Quality Measure #3	Product Quality Measure #4	Product Quality Measure #5
two	low	cust. sat. #1	.64 (.00)	.53 (.00)	n.a.	n.a.	n.a.
		cust. sat. #2	.55 (.00)	.56 (.00)	n.a.	n.a.	n.a.
two	high	cust. sat. #1	.82 (.00)	.05 (.75)	n.a.	n.a.	n.a.
		cust. sat. #2	.74 (.00)	.06 (.77)	n.a.	n.a.	n.a.
five	low	cust. sat. #1	.13 (.35)	.32 (.05)	.18 (.29)	.30 (.07)	.38 (.02)
		cust. sat. #2	.19 (.24)	.37 (.05)	.47 (.03)	.14 (.43)	-.21 (.22)
five	high	cust. sat. #1	.35 (.11)	.71 (.01)	.50 (.03)	-.65 (.01)	-.03 (.91)
		cust. sat. #2	.10 (.71)	.65 (.05)	-.49 (.08)	.21 (.48)	.32 (.31)

TABLE 1 (continued)

PANEL C - Correlation Matrices [3,4]

Two Measures / Low Multicollinearity in Product Quality Measures:

	Prod. Qual. 1	Prod. Qual. 2	Cust. Sat. 1	Cust. Sat. 2
Prod. Qual. 1	1.00			
Prod. Qual. 2	0.10	1.00		
Cust. Sat. 1	0.69 (**)	0.59 (**)	1.00	
Cust. Sat. 2	0.60 (**)	0.61 (**)	0.57 (**)	1.00

Two Measures / High Multicollinearity in Product Quality Measures:

	Prod. Qual. 1	Prod. Qual. 2	Cust. Sat. 1	Cust. Sat. 2
Prod. Qual. 1	1.00			
Prod. Qual. 2	0.67 (**)	1.00		
Cust. Sat. 1	0.86 (**)	0.61 (**)	1.00	
Cust. Sat. 2	0.78 (**)	0.56 (*)	0.48 (*)	1.00

Five Measures / Low Multicollinearity in Product Quality Measures:

	Prod. Qual. 1	Prod. Qual. 2	Prod. Qual. 3	Prod. Qual. 4	Prod. Qual. 5	Cust. Sat. 1	Cust. Sat. 2
Prod. Qual. 1	1.00						
Prod. Qual. 2	0.30	1.00					
Prod. Qual. 3	0.40	0.51 (*)	1.00				
Prod. Qual. 4	0.09	0.34	0.47 (*)	1.00			
Prod. Qual. 5	0.19	0.35	0.06	0.43	1.00		
Cust. Sat. 1	0.39	0.68 (**)	0.56 (*)	0.66 (**)	0.65 (**)	1.00	
Cust. Sat. 2	0.46 (*)	0.64 (**)	0.78 (**)	0.41	0.04	0.41	1.00

Five Measures / High Multicollinearity in Product Quality Measures:

	Prod. Qual. 1	Prod. Qual. 2	Prod. Qual. 3	Prod. Qual. 4	Prod. Qual. 5	Cust. Sat. 1	Cust. Sat. 2
Prod. Qual. 1	1.00						
Prod. Qual. 2	0.69 (**)	1.00					
Prod. Qual. 3	0.70 (**)	0.72 (**)	1.00				
Prod. Qual. 4	0.80 (**)	0.76 (**)	0.73 (**)	1.00			
Prod. Qual. 5	0.67 (**)	0.83 (**)	0.78 (**)	0.68 (**)	1.00		
Cust. Sat. 1	0.66 (**)	0.80 (**)	0.76 (**)	0.53 (*)	0.74 (**)	1.00	
Cust. Sat. 2	0.59 (**)	0.79 (**)	0.44	0.64 (**)	0.68 (**)	0.66 (**)	1.00

* Correlation is significant at the .05 level (two-tailed).
** Correlation is significant at the .01 level (two-tailed).

TABLE 1 (continued)

PANEL D – Collinearity Diagnostics

Data Set		VIF Values		Condition Index
No. of Measures	Multi-collinearity	Largest VIF	Mean VIF	
two	low	1.01	1.01	29.44
two	high	1.83	1.83	31.17
five	low	2.00	1.61	48.88
five	high	4.24	3.74	71.03

NOTES TO TABLE 1:

1. The means of the learning and judgment data sets do not differ ($t = 0.49, p = .63$). The variances of the learning and judgment data sets do not differ ($F = 0.01, p = .92$).

2. Because the standard deviations of all product quality and customer satisfaction measures are identical within any given data set, the standardized and unstandardized regression weights for that data set are equal.

3. *Z*-tests show that the pairwise correlations in the product quality measures for the two measures-low multicollinearity data set are different from those for the two measures-high multicollinearity data set ($p<.02$). The correlations between the product quality and customer satisfaction measures, and between the two customer satisfaction measures do not differ ($p>.20$).

4. *Z*-tests show that the pairwise correlations in the product quality measures for the five measures-low multicollinearity data set are different from those for the five measures-high multicollinearity data set ($p<.10$), except for the correlation between (a) product quality 2 and product quality 3 ($p=.17$), and (b) product quality 4 and product quality 5 ($p=.14$). The correlations between the product quality and customer satisfaction measures, and between the two customer satisfaction measures do not differ ($p>.28$) except for the correlation between (a) product quality 3 and customer satisfaction 2 ($p=.09$), and (b) product quality 5 and customer satisfaction 2 ($p=.02$)

TABLE 2

SAMPLE OF LEARNING DATA PROVIDED TO PARTICIPANTS [1]

	Product Quality Measures					Customer Satisfaction Measures	
Plant No.	Measure #1	Measure #2	Measure #3	Measure #4	Measure #5	Measure #1	Measure #2
2	71.54	77.70	62.23	68.53	71.94	72.74	71.67
4	62.97	72.68	75.07	68.81	66.13	68.40	75.71
7	65.24	55.90	60.68	61.16	61.66	62.11	58.05
8	67.71	58.92	57.18	58.52	60.47	58.12	59.52
10	52.26	59.80	56.21	66.19	76.21	67.98	53.53
11	59.89	69.65	59.06	61.85	58.97	64.81	59.34
12	59.18	68.62	65.90	68.33	66.81	63.51	65.44
16	62.26	61.23	60.54	50.83	55.21	58.21	56.76
19	55.71	57.13	58.89	70.91	66.29	59.80	59.24
20	77.16	66.43	65.30	60.81	72.82	70.34	67.45
21	71.48	67.71	71.50	73.64	58.11	68.92	73.91
22	68.01	64.47	70.60	70.95	64.72	67.32	72.14
24	68.07	68.51	73.18	70.17	64.63	71.41	63.10
27	62.16	67.25	72.84	62.75	63.04	63.60	72.68
29	57.17	66.89	59.30	64.91	64.89	65.20	61.51
30	70.56	67.30	72.50	62.67	70.48	71.79	65.61
33	69.99	56.01	56.28	59.84	66.96	54.59	58.54
36	58.45	63.77	57.16	53.08	59.49	55.04	61.46
39	58.82	53.95	62.33	55.43	48.51	49.56	64.86
40	61.37	56.07	63.26	70.60	62.67	66.56	59.49

NOTE TO TABLE 2:

1. This data is for the five measure-low multicollinearity experimental condition.

TABLE 3

DESCRIPTIVE STATISTICS

PANEL A – Accuracy of Estimated Cue-Criterion Weights (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing and Environmental Regression Models)

Experimental Condition				*N*	Mean	Std. Dev.	Median
Condition No.	Number of Measures	Multi-collinearity	Task Structure				
1	two	low	indicator	13	0.82	0.18	0.83
2	two	low	construct	13	0.77	0.19	0.82
3	two	high	indicator	13	0.81	0.16	0.85
4	two	high	construct	13	0.80	0.14	0.78
5	five	low	indicator	13	0.83	0.13	0.88
6	five	low	construct	12	0.88	0.06	0.86
7	five	high	indicator	12	0.63	0.15	0.59
8	five	high	construct	12	0.64	0.10	0.61

PANEL B – Judgment Consistency (1 – Mean Absolute Difference Between Unstandardized Regression Weights in Policy-Capturing Models of Participant's First Seven and Last Seven Judgments)

Experimental Condition				*N*	Mean	Std. Dev.	Median
Condition No.	Number of Measures	Multi-collinearity	Task Structure				
1	two	low	indicator	13	0.73	0.17	0.71
2	two	low	construct	13	0.84	0.14	0.88
3	two	high	indicator	13	0.56	0.36	0.60
4	two	high	construct	13	0.46	0.36	0.50
5	five	low	indicator	13	0.61	0.32	0.76
6	five	low	construct	12	0.58	0.30	0.68
7	five	high	indicator	12	0.26	0.59	0.42
8	five	high	construct	12	(0.13)	0.79	(0.10)

TABLE 4

HYPOTHESIS 1 RESULTS --
ACCURACY OF ESTIMATED CUE-CRITERION WEIGHTS

PANEL A - Planned Contrasts for H1: Accuracy of Estimated Cue-Criterion Weights (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing and Environmental Models)

Test No.	Contrast[1]	Difference in Means, Left Side of Inequality – Right Side of Inequality	*t*	*p* (one-tailed)
within indicator structure:				
1	2LI > 5LI	(0.01)	0.18	0.431 (wrong direction)
2	2HI > 5HI	0.19	3.33	0.001
3	(5LI-5HI) > (2LI-2HI)	0.20	2.48	0.008
within construct structure:				
4	2LC > 5LC	(0.12)	2.11	0.019 (wrong direction)
5	2HC > 5HC	0.16	2.81	0.003
6	(5LC-5HC) > (2LC-2HC)	0.28	3.48	0.000
indicator versus construct structure:				
7	2LC > 2LI	(0.05)	0.94	0.175 (wrong direction)
8	2HC > 2HI	(0.01)	0.25	0.404 (wrong direction)
9	5LC > 5LI	0.06	0.99	0.161
10	5HC > 5HI	0.01	0.26	0.396
11	(2LI-2HI) > (2LC-2HC)	0.04	0.49	0.313
12	(5LI-5HI) > (5LC-5HC)	(0.04)	0.52	0.302 (wrong direction)
13	(5LI-5HI)-(2LI-2HI) > (5LC-5HC)-(2LC-2HC)	(0.08)	0.71	0.238 (wrong direction)

TABLE 4 (continued)

PANEL B - Results of ANOVA with Accuracy of Estimated Cue-Criterion Weights as the Dependent Variable (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing and Environmental Models)

Source	*d.f.*	*F*	*p*
Number of Measures	1	3.60	0.06
Multicollinearity	1	13.68	0.00
Task Structure	1	0.00	0.98
Number of Measures x Multicollinearity	1	17.23	0.00
Number of Measures x Task Structure	1	1.45	0.23
Multicollinearity x Task Structure	1	0.00	0.98
Number of Measures x Multicollinearity x Task Structure	1	0.49	0.49

NOTE TO TABLE 4:

1. See Table 3, Panel A for means for each experimental condition. The notation for the experimental conditions used in the table of planned contrasts is as follows:

Condition	No. of Measures	Multicollinearity	Task Structure
2LI	two	low	indicator
2LC	two	low	construct
2HI	two	high	indicator
2HC	two	high	construct
5LI	five	low	indicator
5LC	five	low	construct
5HI	five	high	indicator
5HC	five	high	construct

TABLE 5

HYPOTHESIS 2 RESULTS -- JUDGMENT CONSISTENCY

PANEL A - Planned Contrasts for H2: Judgment Consistency (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing Models of Participant's First Seven and Last Seven Judgments)

Test No.	Contrast[1]	Difference in Means, Left Side of Inequality – Right Side of Inequality	*t*	*p* (one-tailed)
within indicator structure:				
1	2LI > 5LI	0.11	0.68	0.250
2	2HI > 5HI	0.31	1.84	0.035
3	(5LI-5HI) > (2LI-2HI)	0.19	0.82	0.206
within construct structure:				
4	2LC > 5LC	0.26	1.55	0.062
5	2HC > 5HC	0.60	3.59	0.000
6	(5LC-5HC) > (2LC-2HC)	0.34	1.44	0.076
indicator versus construct structure:				
7	2LC > 2LI	0.11	0.69	0.245
8	2HC > 2HI	(0.10)	0.60	0.278 (wrong direction)
9	5LC > 5LI	(0.03)	0.18	0.431 (wrong direction)
10	5HC > 5HI	(0.39)	2.34	0.011 (wrong direction)
11	(2LI-2HI) > (2LC-2HC)	(0.21)	0.91	0.182 (wrong direction)
12	(5LI-5HI) > (5LC-5HC)	(0.36)	1.53	0.065 (wrong direction)
13	(5LI-5HI)-(2LI-2HI) > (5LC-5HC)-(2LC-2HC)	(0.15)	0.44	0.332 (wrong direction)

TABLE 5 (continued)

PANEL B - Results of ANOVA with Judgment Consistency as the Dependent Variable (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing Models of Participant's First Seven and Last Seven Judgments)

Source	*d.f.*	*F*	*p*
Number of Measures	1	14.18	0.00
Multicollinearity	1	22.81	0.00
Task Structure	1	1.42	0.24
Number of Measures x Multicollinearity	1	2.48	0.12
Number of Measures x Task Structure	1	1.66	0.20
Multicollinearity x Task Structure	1	2.89	0.09
Number of Measures x Multicollinearity x Task Structure	1	0.19	0.67

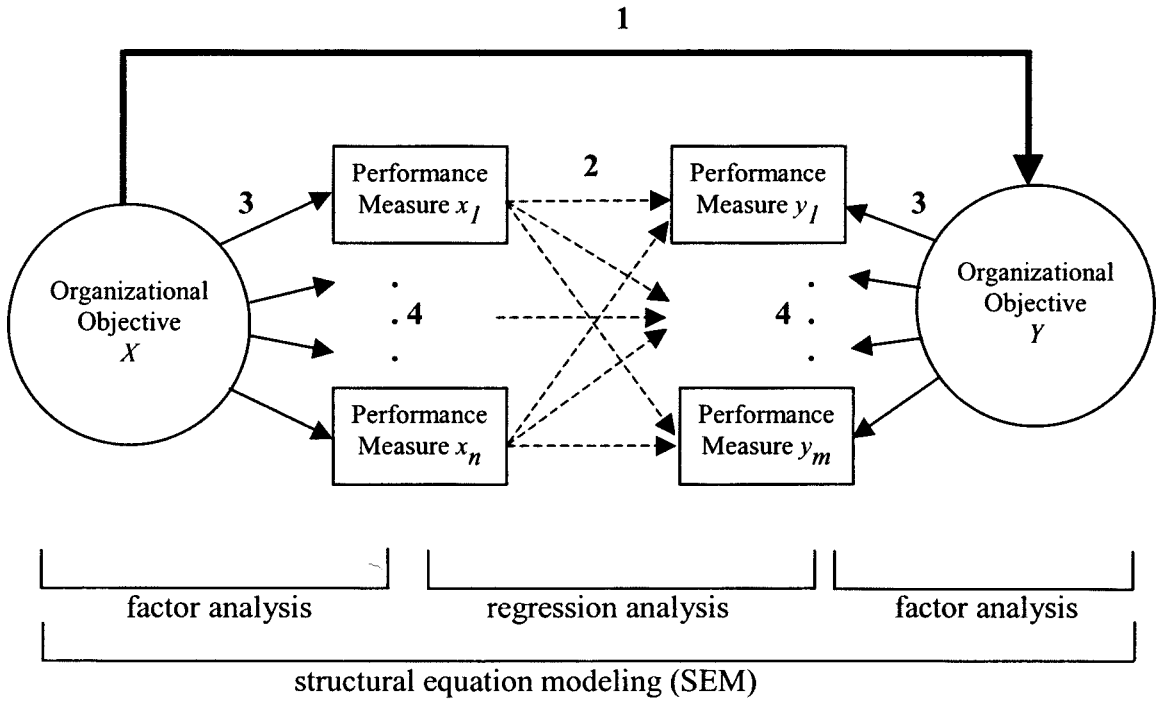
NOTE TO TABLE 5:

1. See Table 3, Panel B for means for each experimental condition. The notation for the experimental conditions used in the table of planned contrasts is as follows:

Condition	No. of Measures	Multicollinearity	Task Structure
2LI	two	low	indicator
2LC	two	low	construct
2HI	two	high	indicator
2HC	two	high	construct
5LI	five	low	indicator
5LC	five	low	construct
5HI	five	high	indicator
5HC	five	high	construct

FIGURE 1

RELATIONS BETWEEN MULTIPLE PERFORMANCE MEASURES



KEY TO FIGURE 1:

◯ Organizational objective (i.e., construct).

▭ Performance measure (i.e., indicator).

→ 1 - Relation between organizational objectives.

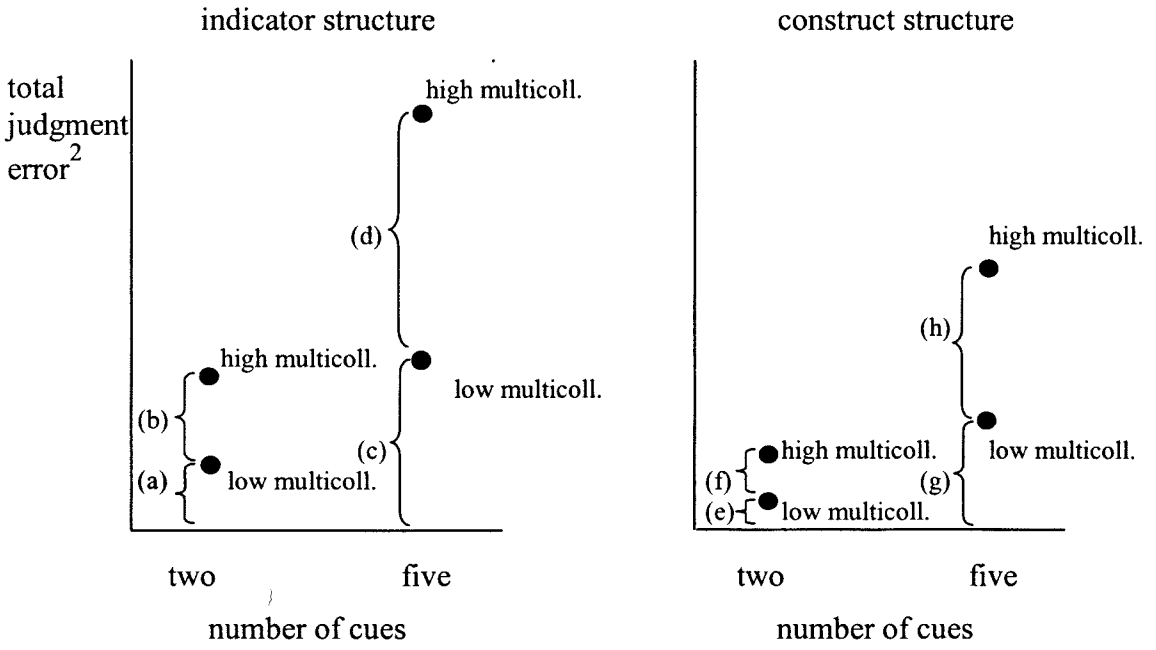
⇢ 2 - Relations between casually-related performance measures.

⟶ 3 - Relations between performance measures and a given organizational objective.

. . . 4 - Relations between performance measures for a given organizational objective (i.e., multicollinearity).

FIGURE 2

SOURCES OF ERROR WITH MULTIPLE MEASURES, MULTICOLLINEARITY AND TASK STRUCTURE [1]



NOTES TO FIGURE 2:

1. Graphs show the relative order of mean total judgment error only.

2.

total judgment error = f		processing error	when multicollinearity is low
	+	attentional error	when multicollinearity is high and ignored
	+	(attentional error x processing error)	when multicollinearity is high and ignored and imperfect adjustments are made

KEY TO FIGURE 2:

(a), (c) Processing error (from estimating weights and applying them to cues) if an indicator structure is used; (c)>(a) since with five cues there are more weights to estimate and apply than there are with two cues and thus total error is expected to be larger.

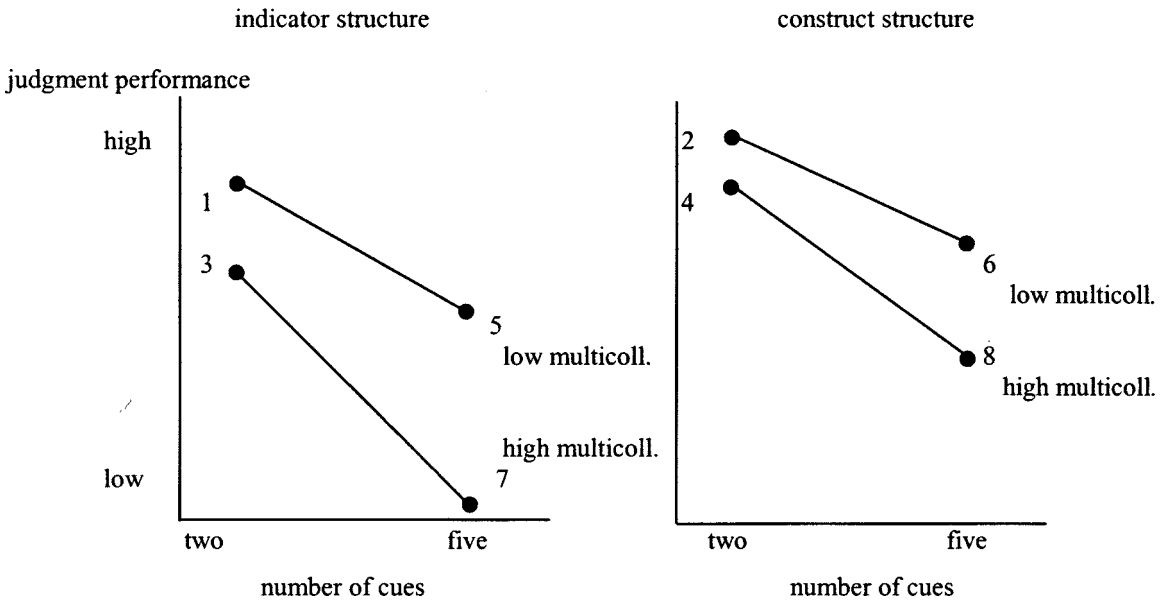
(b), (d) Attentional error (from ignoring high multicollinearity) plus attentional-by-processing error (if high multicollinearity is ignored and imperfect adjustments are made to weights) if an indicator structure is used; (d)>(b) since more cue-cue relations are ignored with five cues (ten relations) than with two cues (one relation) and thus total error is expected to be larger.

(e), (g) Processing error if a construct structure is used; (e)<(a) and (g)<(c) since focusing attention on the lack of multicollinearity via use of the construct structure is expected to result in more careful estimation of cue-criterion weights and lower total error than with use of an indicator structure.

(f), (h) Attentional error plus attentional-by-processing error if a construct structure is used; (f)<(b) and (h)<(d) since focusing attention on multicollinearity via use of the construct structure is expected to result in lower total error than with use of an indicator structure.

FIGURE 3

EXPECTED FORM OF EFFECTS OF
NUMBER OF CUES, MULTICOLLINEARITY, AND TASK STRUCTURE
ON JUDGMENT PERFORMANCE



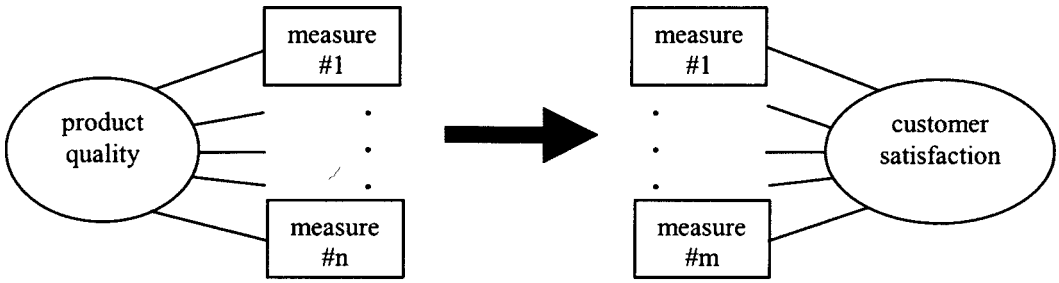
Summary of Predictions:

within indicator	within construct	indicator versus construct
1 > 5	2 > 6	2 > 1
3 > 7	4 > 8	4 > 3
(5-7) > (1-3)	(6-8) > (2-4)	6 > 5
		8 > 7
		(1-3) > (2-4)
		(5-7) > (6-8)
		(5-7)-(1-3) > (6-8)-(2-4)

FIGURE 4

DIAGRAM OF PERFORMANCE MEASUREMENT SYSTEM
PROVIDED TO PARTICIPANTS

Management in your organization wants to learn how product quality affects customer satisfaction. Your organization is implementing a new performance measurement system in which both product quality and customer satisfaction are measured with multiple measures. In visual terms:

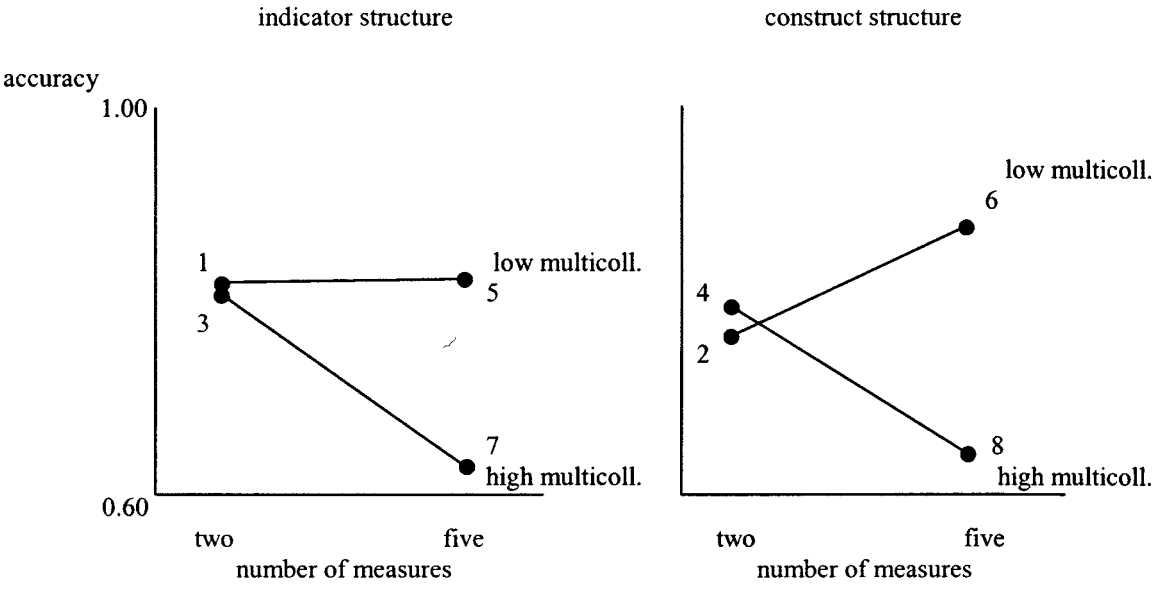


Management is interested in how the particular performance measures they have chosen to use in the new system help you to learn the relation between product quality and customer satisfaction, so they will be asking you to make judgments about how product quality affects customer satisfaction using these measures.

FIGURE 5

HYPOTHESIS 1 RESULTS --
ACCURACY OF ESTIMATED CUE-CRITERION WEIGHTS

H1: Accuracy of Estimated Cue-Criterion Weights (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing and Environmental Models)



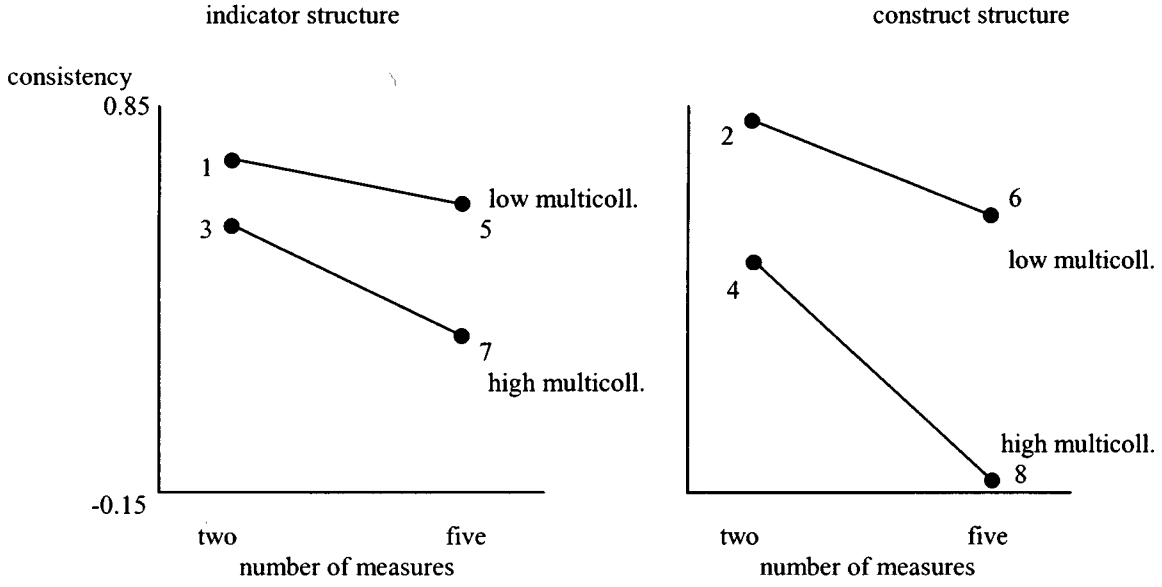
NOTE TO FIGURE 5:

Summary of Predictions (See Table 4, Panel A for Obtained Difference in Means)		*P*-Value (one-tailed) for Contrast
within indicator structure:	1 > 5	0.431 (wrong direction)
	3 > 7	0.001
	(5-7) > (1-3)	0.008
within construct structure:	2 > 6	0.019 (wrong direction)
	4 > 8	0.003
	(6-8) > (2-4)	0.000
indicator versus construct structure:	2 > 1	0.175 (wrong direction)
	4 > 3	0.404 (wrong direction)
	6 > 5	0.161
	8 > 7	0.396
	(1-3) > (2-4)	0.313
	(5-7) > (6-8)	0.302 (wrong direction)
	(5-7)-(1-3) > (6-8)-(2-4)	0.238 (wrong direction)

FIGURE 6

HYPOTHESIS 2 RESULTS --
JUDGMENT CONSISTENCY

H2: Judgment Consistency (1 – Mean Absolute Difference Between Unstandardized Weights in Policy-Capturing Models of Participant's First Seven and Last Seven Judgments)



NOTE TO FIGURE 6:

	Summary of Predictions (See Table 5, Panel A for Obtained Difference in Means)	*P*-Value (one-tailed) for Contrast
within indicator structure:	1 > 5	0.250
	3 > 7	0.035
	(5-7) > (1-3)	0.206
within construct structure:	2 > 6	0.062
	4 > 8	0.000
	(6-8) > (2-4)	0.076
indicator versus construct structure:	2 > 1	0.245
	4 > 3	0.278 (wrong direction)
	6 > 5	0.431 (wrong direction)
	8 > 7	0.011 (wrong direction)
	(1-3) > (2-4)	0.182 (wrong direction)
	(5-7) > (6-8)	0.065 (wrong direction)
	(5-7)-(1-3) > (6-8)-(2-4)	0.332 (wrong direction)

ENDNOTES

1. Literature on information load investigates judgment performance across sets of information with different predictive ability, or the change in judgment performance as measures are incrementally added to a set. These studies often try to determine the point at which individuals reach information overload (Casey 1980; Shields 1980, 1983; Iselin 1988; Chewning and Harrell 1990; Tuttle and Burton 1999). While these settings are different from that examined here, the results provide support for the notion that increases in cognitive processing will result in lower judgment performance.

2. Naylor and Schenck (1968) found that judgment performance increases with multicollinearity in the cues, but their measures are correlational measures of performance which are inflated by multicollinearity (Libby 1981; Ashton 1982).

3. While this dissertation examines how well individuals estimate cue-criterion weights and not how well they make predictions, some prior research suggests that an equal-weight heuristic yields predictions that are not significantly less accurate as are predictions based on OLS weights in a regression model (Dawes and Corrigan 1974; Dawes 1979; Wainer 1976). However, this has been disputed by others, who contend that predictions generated by equal-weight and OLS models are equally as accurate only for a limited range of situations, and in particular the predictions are not equally as accurate when the difference between the highest and lowest OLS regression weights is greater than 0.5, which happens more often when the predictors are more highly correlated (Wainer 1976; Laughlin 1978; Pruzek and Frederick 1978; Barron 1988). See Appendix B for further discussion of how multicollinearity affects OLS regression weights.

4. The discussion that follows assumes that all cue-cue correlations are positive. If all cue-cue correlations are negative and individuals ignore multicollinearity when making estimates for each cue-criterion weight, then their estimated weights will instead be too low, but the same effects on judgment performance are expected to occur as are expected with positive cue-cue correlations. If the cue-cue correlations are mixed in sign (i.e., some are positive, some are negative) and individuals ignore multicollinearity, then it is mathematically possible that errors resulting from using estimated weights that are too high will be offset by those resulting from using estimated weights that are too low. However, prior research has found that individuals have more difficulty cognitively processing negative correlations, so mixed cue-cue correlations are likely to introduce other types of errors that are not examined in this dissertation (Naylor and Clark 1968; Brehmer 1971, 1973b, 1974a; Brehmer, Kuylenstierna and Liljergren 1974).

5. Contrary to this literature, Jiambalvo and Waller (1984) found that decomposition of an audit task did not lead individuals to make different judgments than they did when the task was not decomposed. They attribute this finding to either a failure of the decomposition to direct attention to the critical parts of the task or a failure

of individuals to process information even when the decomposition focused attention on it.

6. Cohen and Cohen (1983, p. 161) define a "medium" effect in behavioral sciences research an R^2 of 0.13 (which translates to an f^2 of 0.15). If the average adjusted-R^2 that was obtained in the ANOVA's for the dependent variables had been used as the effect size in the power analysis, then results would have indicated that a sample size of 63 would provide power of 90%.

7. The transformation applied to the values of the product quality and customer satisfaction measures produced by the data generation program to obtain the values for the learning data set was:

target mean +
[(current value of variable - current mean)
x (target standard deviation / current standard deviation)].

8. In the third of five experiments, Broniarczyk and Alba (1994) found that judgment performance was lower when individuals examined all cases at once in a tabular format rather than when they examined them case-by-case and made predictions after each case. However, the tabular format was presented on paper, while the case-by-case format was presented on a computer screen. Therefore, it is not possible to determine whether the lower performance was due to the format in which the information was presented (tabular versus case-by-case), the requirement (or lack of) to make case-by-case judgments, or the paper versus pencil presentation.

9. The transformation applied to the values of the product quality and customer satisfaction measures in the learning data sets to obtain the values for the judgment data sets was the same as the formula in Endnote 7.

10. This statement was verified in personal communications with Dr. Connie Page (Professor of Statistics and Probability and director of the Statistical Consulting Service), and Dr. Alexander Von Eye (Professor Psychology and author or editor of eight statistics textbooks for research in the social sciences), both at Michigan State University.

11. Both dependent variables used in tests are computed using the participants' judgments of customer satisfaction measure #1. The dependent variables were also computed using participants' judgments of customer satisfaction measure #2. See Endnote 15 for results of hypothesis tests for dependent variables computed using participants' judgments of customer satisfaction measure #2.

12. Besides multicollinearity, it is possible that there could be other factors that might result in inequality of the error variances across the experimental conditions. However, since the results of Levene's test and tests performed using the

transformed judgment consistency dependent variable indicate that inequality of error variances is not a significant statistical problem in the analyses, other potential sources of error variance are not investigated.

13. The error terms for the dependent measure of the accuracy of estimated OLS weights in the policy-capturing model were normally distributed (K-S test, $p>.05$). The error terms for both the dependent measure of judgment consistency and the $\log_{10}$ transformation of the dependent measure of judgment consistency were normally distributed ($p>.05$).

14. The number of measures (two or five) and multicollinearity (low or high) independent variables were manipulations of the data participants used for the judgment task, not manipulations about participants' beliefs or knowledge. Therefore, no manipulation checks were necessary for these independent variables. Similarly, the task structure (indicator or construct) independent variable was a manipulation of the types of judgments participants made to complete the task. All participants made judgments in accordance with the task structure to which they were assigned, so no further manipulation checks were necessary.

15. When the dependent variables were computed using the participants' judgments of customer satisfaction measure #2, the results of ANOVA's and the pattern of means were comparable to those for customer satisfaction measure #1, except for lower accuracy in the two-measure-high-multicollinearity conditions for judgments of customer satisfaction measure #2 than customer satisfaction measure #1. While the environmental regression model for customer satisfaction measure #2 has a lower adjusted-R^2 than that for customer satisfaction measure #1 (Table 1, Panel A), all other features of the data sets are the same (see Chapter 3, "Experimental Setting" subsection), and the lower adjusted-R^2 did not affect accuracy as much in the other experimental conditions as it did in the two-measure-high-multicollinearity condition. Further, the demands of the task were similar across experimental conditions and randomization of participants appeared to be successful (see Chapter 4, "Tests of Randomization and Sensitivity of Results" subsection). Therefore, it does not appear that the lower accuracy in the two-measure-high-multicollinearity conditions is due to differences in the data sets used in the task, fatigue, or differences in participants. While this accuracy result for customer satisfaction measure #2 was not anticipated, it suggests that under certain conditions, participants may also have difficulty accurately estimating cue-criterion weights with two measures, but what those conditions are were not measured in this study. The remainder of the results in this dissertation are reported using the dependent variables computed for judgments of customer satisfaction measure #1, but it is possible that the impact of multicollinearity on accuracy may be understated when there are two measures.

APPENDICES

APPENDIX A

SMALL-SCALE EMPIRICAL INVESTIGATION OF HEURISTICS USED IN PREDICTIVE JUDGMENT TASK

Hutchinson and Alba (1997) note that few studies have investigated how individuals estimate relations between numeric variables, especially with respect to estimating cross-sectional correlations which are then used as a component of judgments, as is the case in this dissertation. Many studies investigate how cues affect judgments (i.e., input-output effects) rather than the process of how people estimate the relations between cues and a criterion that they then use to make judgments (i.e., process effects). Hutchinson and Alba (1997) investigate heuristics used in covariation assessment with numeric variables across different contexts, but three of their four experiments use time-series data, which can prompt individuals to use very different heuristics than they might use with cross-sectional data. In their experiment that did use cross-sectional data, the judgment task differed from that used here. Because prior research on cognitive heuristics that are applied to the predictive judgment task in this dissertation is limited, a small-scale empirical investigation of how individuals do this task was conducted.

A convenience sample of seven individuals were asked to orally and concurrently describe their approach to a judgment task similar to the one in this dissertation (although smaller in scale) while doing the task. Over the course of two days, the individuals reported to a room one at a time. Upon arriving, I described to each person that I was interested in understanding how he or she would approach a common business judgment task, and that there was no right or wrong way to approach or complete the task. I told him or her that I would be writing down the steps he or she followed to complete various versions of the same task, and that he or she was to "think out loud" as they did so. Once the participant understood what would happen, he or she was given paper-and-pencil versions of the task materials. I manually wrote down what he or she said during task

execution, but did not interject in any way or prompt him or her to describe particular heuristics or processes. Immediately after each individual completed the task and left the room, I reviewed the transcript from his or her session and coded it in accordance with descriptions of heuristics in Hutchinson and Alba (1997). The only heuristic used by any of the seven individuals that was not described in Hutchinson and Alba (1997) was an equal-weight heuristic, but use of this heuristic has been documented in other research and was thus easily coded (Peterson et al. 1965; Brehmer 1973a; Nisbett et al. 1981; Bloomfield et al. 1998a, b). None of the individuals used heuristics or steps in task execution that were otherwise unidentifiable.

Two of the seven individuals were Ph.D. students with extensive statistical training, one was a senior manager in a Big Five firm, and four were upper-level undergraduate students. The number of cues was manipulated within subjects (i.e., each individual concurrently described how they did both a two-cue and for a five-cue judgment task). The multicollinearity in the cues was manipulated between subjects (three individuals had cues with low multicollinearity and four individuals had cues with high multicollinearity).

Regardless of the number of cues, all seven individuals began each prediction task by estimating weights to place on the cues in a manner consistent with either a difference heuristic or an equal-weight heuristic in which the weights are based on the inverse of the number of cues in the task (i.e., 0.5 and 0.2 for two and five cues, respectively). This provides support for the assumption that the number of cues does not affect the heuristic individuals use to estimate cue-criterion weights (i.e., there is no between-task switching

of heuristics), and that individuals' focus is on the bivariate cue-criterion relations and not on multicollinearity.

Many of the individuals in Hutchinson and Alba's (1997) study used an "exemplar-based" heuristic, which when applied to this task would involve comparing one set of potential cue values to the series of past cue-criterion observations to find the best match, and using the criterion value from that match as the judgment. None of the individuals in this investigation used this heuristic to generate their judgments. As noted in the following paragraph, however, this heuristic was employed to check the reasonableness of judgments.

After the individuals concurrently described how they estimated cue-criterion weights, they concurrently described how they applied the weights to a series of cue values in the same data set. Six included a step in which they checked the resulting judgments to observed values of the criterion in the same data set. If individuals believed their judgments differed too much from the criterion values to which they were compared, then they switched to the other heuristic to estimate the weights (i.e., there was within-task switching of heuristics). If they again checked judgments based on the second heuristic back to the observed criterion values and believed their judgments differed too much from those values, then they tried to think of a different heuristic that could be applied to the task. Two individuals said they would like to draw multiple x-y plots to give them a feel of potential relations but conceded it was virtually impossible to do. The remaining did not articulate any other distinct heuristics they tried.

In all cases, if individuals believed their predictions were too far from observed criterion values after use of their second heuristic, then they adjusted their estimations.

The adjustments were sometimes based on deducting or adding some numeric or percentage amount from the prediction or the weights, but the individuals usually described these adjustments as 'winging it', 'eyeballing the data', or 'guesstimating'. When there was high multicollinearity in the cues, the magnitude of the adjustments the individuals wanted to make made it more difficult for them to explain what those adjustments entailed or a systematic method they were using to estimate those adjustments. This provides further support for the assumption that individuals switch heuristics within a task, and support for the assumption that individuals make imperfect adjustments to either their cue weights or their judgments.

One of the Ph.D. students suggested that multicollinearity in the cues could drive differences between his judgments and the observed criterion values, but was unsure of how to integrate that into his judgments. None of the other individuals mentioned multicollinearity during the tasks. This provides further support that individuals tend to focus on bivariate cue-criterion relations when doing this task.

The two Ph.D. students approached the judgment tasks in the same way as the other individuals. Therefore, it does not appear that greater statistics knowledge had an impact on heuristic choice or use.

APPENDIX B

ILLUSTRATION OF COMPUTATION OF REGRESSION WEIGHTS FOR MODELS WITH TWO OR FIVE INDEPENDENT VARIABLES

To estimate the frequency with which a regression weight for an independent variable is negative and significant ($p<.20$) although all values in $\mathbf{r}_{XX}$ and $\mathbf{r}_{YX}$ are positive, 100 data sets (50 with $n=20$, 50 with $n=50$) with five independent variables and high multicollinearity were generated. The occurrences of negative weights (significant at $p<.20$) on independent variables with positive correlations to the dependent variable is shown in the table that follows.

TABLE A1

OCCURENCES OF NEGATIVE REGRESSION WEIGHTS IN DATA SETS WITH HIGH MULTICOLLINEARITY

	$n = 20$		$n = 50$	
Number of Significant Negative Weights ($p<.20$) in Regression Model	Occurred in:		Occurred in:	
	Number of Data Sets	Percent of Data Sets	Number of Data Sets	Percent of Data Sets
0	6	12 %	0	--
1	31	62 %	12	24 %
2	13	26 %	36	72 %
3	--	--	2	4 %
Total	50	100 %	50	100 %

Negative regression weights can result with high multicollinearity since the variance of the estimated regression weights increases (see Appendix C). The following is a matrix algebra example of why this can occur, based on the formulas for standardized regression weights.

The formula for computing standardized regression weight is: [1]

$$\mathbf{b} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

where $\mathbf{b}$ = vector of standardized regression weights,
$\mathbf{r}_{XX}$ = matrix of simple correlations between pairs of x_i measures, and
$\mathbf{r}_{YX}$ = vector of simple correlations between y and each x_i measure.

For a regression model with two independent variables, the b_i weights are computed as follows:

$$b_1 = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2}$$

$$b_2 = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}$$

where b_i = regression weight for x_i,
r_{Yi} = correlation between y and x_i, and
r_{12} = correlation between x_1 and x_2.

As can be seen from this formula, when $r_{12} = 0$, no adjustment is needed for multicollinearity, and the standardized regression weights for x_i are equal to the bivariate or zero-order correlation between x_i and y.

For a regression model with five independent variables, the formula for the standardized regression weights as written in matrix form is:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & 1 & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & 1 & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & 1 & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \\ r_{y4} \\ r_{y5} \end{bmatrix}.$$

Thus, the formula for each standardized regression weight is:

$$
\begin{aligned}
b_1 &= 1(r_{y1}) + 0(r_{y2}) + 0(r_{y3}) + 0(r_{y4}) + 0(r_{y5}) &&= r_{y1} \\
b_2 &= 0(r_{y1}) + 1(r_{y2}) + 0(r_{y3}) + 0(r_{y4}) + 0(r_{y5}) &&= r_{y2} \\
b_3 &= 0(r_{y1}) + 0(r_{y2}) + 1(r_{y3}) + 0(r_{y4}) + 0(r_{y5}) &&= r_{y3} \\
b_4 &= 0(r_{y1}) + 0(r_{y2}) + 0(r_{y3}) + 1(r_{y4}) + 0(r_{y5}) &&= r_{y4} \\
b_5 &= 0(r_{y1}) + 0(r_{y2}) + 0(r_{y3}) + 0(r_{y4}) + 1(r_{y5}) &&= r_{y5}\,.
\end{aligned}
$$

In the case of zero correlations between the independent variables, $\mathbf{r}_{XX}$ is the identity matrix. Since the inverse of the identity matrix is the identity matrix, the equation above becomes:

$$
\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{y1} \\ r_{y1} \\ r_{y3} \\ r_{y4} \\ r_{y5} \end{bmatrix}.
$$

As is the case with two independent variables, the regression weight for each independent variable is equal to its bivariate or zero-order correlation with the dependent variable.

As the number of variables and multicollinearity increases, it is possible to obtain negative regression weights, although $\mathbf{r}_{XX}$ and $\mathbf{r}_{YX}$ contain only positive elements.[2] The variance in the regression weights increases as multicollinearity increases (Greene 2000, p. 257; Kennedy 2001, p. 184-185), so as the number of measures and multicollinearity increase, the potential range of values for the weights increases. In matrix algebra form, holding the off-diagonal elements of $\mathbf{r}_{XX}$ constant, for a given row i in $\mathbf{r}_{XX}^{-1}$ the relative weight of r_{yi} in computing b_i is reduced. Depending on the magnitude of the off-diagonal elements in $\mathbf{r}_{XX}$ and the elements in $\mathbf{r}_{YX}$, a small r_{yi} can

be offset by larger correlations between the other independent variables themselves and between the other independent variables and the dependent variable. This is best illustrated by a simplified example using hypothetical correlation matrices.

Given: The formula for standardized regression weights of $\mathbf{b} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$

Case 1 - Two Independent Variables and Low Values for Off-diagonal Elements of $\mathbf{r}_{XX}$

Suppose that there are two independent variables and low multicollinearity is represented by an $\mathbf{r}_{XX}$ matrix with the values of 0.15 on the off-diagonals:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.15 \\ 0.15 & 1.00 \end{bmatrix}.$$

Also suppose that:

$$\mathbf{r}_{YX} = \begin{bmatrix} 0.60 \\ 0.90 \end{bmatrix}.$$

The inverse of $\mathbf{r}_{XX}$ is:

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 1.02 & -0.15 \\ -0.15 & 1.02 \end{bmatrix}.$$

The formulas to compute each element of **b** are:

$$b_1 = 1.02(0.60) - 0.15(0.90) = 0.48$$
$$b_2 = -0.15(0.60) + 1.02(0.90) = 0.83 .$$

Case 2 - Two Independent Variables and High Values for Off-diagonal Elements of $\mathbf{r}_{XX}$

Suppose that there are two independent variables and high multicollinearity is represented by an $\mathbf{r}_{XX}$ matrix with the values of 0.85 on the off-diagonals:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.85 \\ 0.85 & 1.00 \end{bmatrix}.$$

Like Case 1, also suppose that:

$$\mathbf{r}_{YX} = \begin{bmatrix} 0.60 \\ 0.90 \end{bmatrix}.$$

The inverse of $\mathbf{r}_{XX}$ is:

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 3.60 & -3.06 \\ -3.06 & 3.60 \end{bmatrix}.$$

The formulas to compute each element of **b** are:

$$b_1 = 3.60(0.60) - 3.06(0.90) = -0.59$$
$$b_2 = -3.06(0.60) + 3.60(0.90) = 1.41 .$$

Case 3 - Five Independent Variables and Low Values for Off-diagonal Elements of $\mathbf{r}_{XX}$

Suppose that there are five independent variables and low multicollinearity is represented by an $\mathbf{r}_{XX}$ matrix with the values of 0.15 on the off-diagonals:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.15 & 0.15 & 0.15 & 0.15 \\ 0.15 & 1.00 & 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 1.00 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 & 1.00 & 0.15 \\ 0.15 & 0.15 & 0.15 & 0.15 & 1.00 \end{bmatrix}.$$

Also suppose that:

$$\mathbf{r}_{YX} = \begin{bmatrix} 0.50 \\ 0.60 \\ 0.70 \\ 0.80 \\ 0.90 \end{bmatrix}.$$

The inverse of $\mathbf{r}_{XX}$ is:

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 1.07 & -0.11 & -0.11 & -0.11 & -0.11 \\ -0.11 & 1.07 & -0.11 & -0.11 & -0.11 \\ -0.11 & -0.11 & 1.07 & -0.11 & -0.11 \\ -0.11 & -0.11 & -0.11 & 1.07 & -0.11 \\ 0\text{-}.11 & -0.11 & -0.11 & -0.11 & 1.07 \end{bmatrix}.$$

The formulas to compute each element of **b** are:

$$\begin{aligned} b_1 &= 1.07(0.50) - 0.11(0.60) - 0.11(0.70) - 0.11(0.80) - 0.11(0.90) = 0.20 \\ b_2 &= -0.11(0.50) + 1.07(0.60) - 0.11(0.70) - 0.11(0.80) - 0.11(0.90) = 0.32 \\ b_3 &= -0.11(0.50) - 0.11(0.60) + 1.07(0.70) - 0.11(0.80) - 0.11(0.90) = 0.44 \\ b_4 &= -0.11(0.50) - 0.11(0.60) - 0.11(0.70) + 1.07(0.80) - 0.11(0.90) = 0.56 \\ b_5 &= -0.11(0.50) - 0.11(0.60) - 0.11(0.70) - 0.11(0.80) + 1.07(0.90) = 0.67 . \end{aligned}$$

Case 4 - Five Independent Variables and High Values for Off-diagonal Elements of $\mathbf{r}_{XX}$

Suppose that there are five independent variables and high multicollinearity is represented by an $\mathbf{r}_{XX}$ matrix with the values of 0.85 on the off-diagonals:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.85 & 0.85 & 0.85 & 0.85 \\ 0.85 & 1.00 & 0.85 & 0.85 & 0.85 \\ 0.85 & 0.85 & 1.00 & 0.85 & 0.85 \\ 0.85 & 0.85 & 0.85 & 1.00 & 0.85 \\ 0.85 & 0.85 & 0.85 & 0.85 & 1.00 \end{bmatrix}.$$

Like Case 3, also suppose that:

$$\mathbf{r}_{YX} = \begin{bmatrix} 0.50 \\ 0.60 \\ 0.70 \\ 0.80 \\ 0.90 \end{bmatrix}.$$

The inverse of $\mathbf{r}_{XX}$ is:

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 5.38 & -1.29 & -1.29 & -1.29 & -1.29 \\ -1.29 & 5.38 & -1.29 & -1.29 & -1.29 \\ -1.29 & -1.29 & 5.38 & -1.29 & -1.29 \\ -1.29 & -1.29 & -1.29 & 5.38 & -1.29 \\ -1.29 & -1.29 & -1.29 & -1.29 & 5.38 \end{bmatrix}.$$

The formulas to compute each element of **b** are:

$$\begin{aligned} b_1 &= 5.38(0.50) - 1.29(0.60) - 1.29(0.70) - 1.29(0.80) - 1.29(0.90) = -1.17 \\ b_2 &= -1.29(0.50) + 5.38(0.60) - 1.29(0.70) - 1.29(0.80) - 1.29(0.90) = -0.51 \\ b_3 &= -1.29(0.50) - 1.29(0.60) + 5.38(0.70) - 1.29(0.80) - 1.29(0.90) = 0.16 \\ b_4 &= -1.29(0.50) - 1.29(0.60) - 1.29(0.70) + 5.38(0.80) - 1.29(0.90) = 0.83 \\ b_5 &= -1.29(0.50) - 1.29(0.60) - 1.29(0.70) - 1.29(0.80) + 5.38\,(0.90) = 1.49\,. \end{aligned}$$

In the low multicollinearity cases (Cases 1 and 3), the off-diagonal elements of $\mathbf{r}_{XX}^{-1}$ have less of an impact on the computation of b_i than they do in the high multicollinearity cases (Cases 2 and 4).

The following table shows that when the off-diagonal elements of $\mathbf{r}_{XX}$ are held constant, the relative weight of r_{yi} in computing the value for a given b_i is smaller. The values indicate by how much the weight on r_{yi} exceeds the sum of the weights on the remaining independent-dependent variable correlations in $\mathbf{r}_{YX}$ when computing **b** (i.e., the sum of the elements of any row of $\mathbf{r}_{XX}^{-1}$).

TABLE A2

RELATIVE WEIGHTS OF r_{yi} IN COMPUTING b_i
WITH DIFFERING DEGREES OF MULTICOLLINEARITY

	For a Given Row of $\mathbf{r}_{XX}^{-1}$, Difference Between Value of b_i and Sum of Values for All b_j	
Value of Off-Diagonal Elements of $\mathbf{r}_{XX}$	Two Independent Variables	Five Independent Variables
0.05	0.95	0.83
0.10	0.91	0.71
0.15	0.87	0.63
0.20	0.83	0.56
0.25	0.80	0.50
0.30	0.77	0.45
0.35	0.74	0.42
0.40	0.71	0.38
0.45	0.69	0.36
0.50	0.67	0.33
0.55	0.65	0.31
0.60	0.63	0.29
0.65	0.61	0.28
0.70	0.59	0.26
0.75	0.57	0.25
0.80	0.56	0.24
0.85	0.54	0.23
0.90	0.53	0.22
0.95	0.51	0.21

NOTES TO APPENDIX B:

1. Neter et al. 1996, pp. 279-282.

2. This was verified in personal communications with Dr. Connie Page (Professor of Statistics and Probability and director of the Statistical Consulting Service), and Dr. Alexander Von Eye (Professor Psychology and author or editor of eight statistics textbooks for research in the social sciences), both at Michigan State University.

APPENDIX C

NOTES ON EFFECTS OF MULTICOLLINEARITY ON ESTIMATED WEIGHTS IN OLS REGRESSION

As multicollinearity in the independent variables used in an OLS regression model increases, the variances of the estimated regression weights become larger, and thus the weights are less precise (Greene 2000, p. 257; Kennedy 2001, p. 184-185). Therefore, it is possible that multicollinearity can be unduly influencing computations of the dependent variables used in this dissertation and thus the hypothesis tests.

OLS remains the best linear unbiased estimator even in the presence of high multicollinearity (Gujarati 1988, p. 288; Greene 2000, p. 256; Kennedy 2001, p. 184), so bias is not a concern. However, multicollinearity does raise other issues.

The dependent measures for accuracy of estimated cue-criterion weights and judgment consistency are computed using weights from OLS regression models. Since variances of regression weights increase with high multicollinearity (i.e., there is likely to be more measurement error in the dependent variables for the high multicollinearity experimental conditions), it is possible that the homogeneity of variance assumption of OLS is violated in hypothesis tests using these measures. However, a review of the statistics literature shows that higher variances in regression weights in the high multicollinearity conditions do not severely restrict the ability to make inferences in statistical tests using these measures (Gujarati 1988; Neter et al. 1996; Von Eye and Schuster 1998; Greene 2000; Kennedy 2001):

1) Based on Levene's test, the variances of the error terms did not differ for the measure of accuracy of estimated cue-criterion weights ($p>.05$), but differed for the measure of judgment consistency ($p<.05$). A $\log_{10}$ transformation of the judgment consistency measure eliminated this violation, but results of hypothesis tests using the transformed variable were not qualitatively different than those

using the original variable. Further, Neter et al. (1996, p. 776), Von Eye and Schuster (1998, p. 179), and Greene (2000, p. 501) note OLS regression and *F*-tests are robust to violations of the homogeneity of variance assumption, which is consistent with the fact that results of tests using transformed and non-transformed measures did not differ. Therefore, these violations should not have a significant impact on the interpretation of the results.

2) One rule-of-thumb to judge whether multicollinearity is unduly influencing the estimates of regression weights is that if VIF values are 10 or larger, then the estimated standardized regression weights may be unduly influenced by the multicollinearity (Neter et al. 1996, p. 387; Von Eye and Schuster 1998, p. 137; Kennedy 2001, p. 190). The largest of the VIF measures for the data sets used in this dissertation is 4.24 (see Table 3, Panel D).

3) Another rule-of-thumb to judge whether multicollinearity is unduly influencing the estimates of regression weights is when the condition index is greater than 20 (Greene 2000, p. 258) or 30 (Gujarati 1988, p. 301; Kennedy 2001, p. 190). Of the data sets used in this dissertation, all of the condition indices exceed 20, which is Greene's (2000) cutoff, while the condition indices for all but the two measure/low multicollinearity data set exceed 30, which is Gujarati's (1988) and Kennedy's (2001) cutoff (see Table 3, Panel D). While this index does seem to indicate that the estimates of regression weights may be unduly influenced by multicollinearity, Gujarati (1988, p. 302) notes that the view that this is the best multicollinearity diagnostic is not widely shared.

4) Several texts indicate that in the presence of multicollinearity, estimates of regression weights change dramatically with even slight changes in the data matrix (Gujarati 1988, p. 294; Neter et al. 1996, p. 385; Greene 2000, p. 256; Kennedy 2001, pp. 189-190). To test if that is the case in the data sets used in this dissertation, a random variable from a uniform distribution with a range of two standard deviations was added to each product quality measure. The new values of the product quality measures were then used in regressions of the product quality measures on customer satisfaction. The regression weights in models using product quality measures which had been changed did not dramatically differ in sign or relative magnitude from the weights for the unchanged product quality measures.

5) Another rule-of-thumb suggests that if the adjusted-R^2 of any regression model of one independent measure on the other independent measures exceeds the adjusted-R^2 of the full model, then multicollinearity may be severely influencing the estimates of the regression weights (Greene 2000, p. 258; Kennedy 2001, p. 187). That was not the case for any of the data sets used in this dissertation.

APPENDIX D

EXPERIMENTAL MATERIALS

Envelope 1
Pre-Experiment Questionnaire

Materials Are the Same Across All Experimental Conditions

Assume that an organization manufactures a moderately priced consumer product (not a car, but not a candy bar, either), and you are interested in how product quality affects customer satisfaction with the product. Both product quality and customer satisfaction are measured using 0-100 scales, with 0 being the lowest possible level (i.e., lowest product quality, lowest customer satisfaction) and 100 being the highest possible level (i.e., highest product quality, highest customer satisfaction).

Although the effect of product quality on customer satisfaction varies across organizations and products, you probably have some general expectations about the impact of product quality on customer satisfaction based on your past experiences, training, stories in the business press, etc. Even though your expectations are uncertain, they influence what you are willing to believe.

For example, you might not expect customer satisfaction to *decrease* by an extremely large amount (e.g., a 100% decrease) if the product quality measure decreases by a small amount (e.g., 1%), and at the other extreme you might not expect customer satisfaction to *increase* by an extremely large amount (e.g., a 100% increase) for a small increase (e.g., 1%) in product quality. For a 1% change in product quality, think about what range of possible changes in customer satisfaction you believe are likely (your range can include positive and negative numbers and zero).

For a **1% increase** in the product quality measure, by what percentage would you expect the customer satisfaction measure to change? Be sure to specify whether your expected change is positive, negative, or zero. Your change does not have to be in a whole percentage.

For a **1% increase** in product quality, I expect that:

a) the **minimum** change to the customer satisfaction measure will be _____ % (indicate + or -)

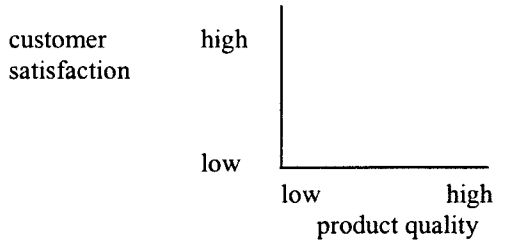
b) the **maximum** change to the customer satisfaction measure will be _____ % (indicate + or -)

c) the **most likely** change to the customer satisfaction measure will be _____ % (indicate + or -)

How confident are you that your expectations are accurate?

1	2	3	4	5	6	7	8	9	10
No Confidence									Extremely High Confidence

On the graph below, please draw what you believe the general relationship between product quality and customer satisfaction looks like (e.g., ／, ⌒, ＼, ⌣).



Assume you are making a purchase of a moderately priced consumer product (not a car, but not a candy bar, either) for your own personal use. Please answer the following questions from your personal perspective.

a) How important is product quality to you when you purchase such a product?

1	2	3	4	5	6	7	8	9	10
Not At All Important									Extremely Important

b) How upset are you when a product you have purchased fails to meet your quality expectations?

1	2	3	4	5	6	7	8	9	10
Not At All Upset									Extremely Upset

c) How likely are you to contact the organization from which you purchased the product to complain if it does not meet your quality expectations?

1	2	3	4	5	6	7	8	9	10
Not At All Likely									Extremely Likely

d) How important is it that your quality expectations are met when you buy a product like this?

1	2	3	4	5	6	7	8	9	10
Not At All Important									Extremely Important

Again, assume an organization manufactures a moderately priced consumer product. In the U.S. economy, how important do you believe the following are to such an organization's long-term financial success?

a) product quality

1	2	3	4	5	6	7	8	9	10
Not At All Important									Extremely Important

b) customer satisfaction

1	2	3	4	5	6	7	8	9	10
Not At All Important									Extremely Important

WHEN FINISHED, RETURN THESE TWO PAGES TO THEIR ORIGINAL ENVELOPE, THEN PROCEED TO THE MATERIALS IN THE NEXT ENVELOPE.

Envelope 2
Learning Materials and Data

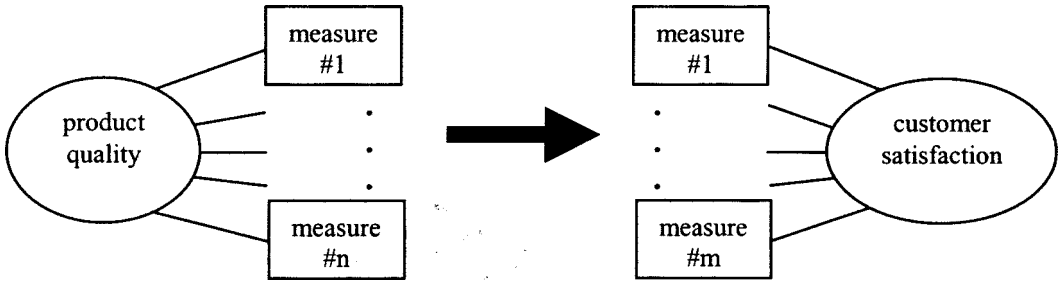
"Your Task" Section Differs for Indicator and Construct Structure Conditions As Noted on the Following Pages

Data on Learning Table Differs by Experimental Condition in Terms of:

1) **Number of Measures of Product Quality (2 or 5)**
2) **Multicollinearity in Product Quality Measures (Low or High)**

Introduction

Management in your organization wants to learn how product quality affects customer satisfaction. Your organization is implementing a new performance measurement system in which both product quality and customer satisfaction are measured with multiple measures. In visual terms:



Management is interested in how the particular performance measures they have chosen to use in the new system help you to learn the relation between product quality and customer satisfaction, so they will be asking you to make judgments about how product quality affects customer satisfaction using these measures.

Information About the Performance Measure Data You Will be Using

The performance measure data that you will use to make these judgments will be labeled "product quality measure #1", "customer satisfaction measure #1", etc. and will be scale-free, which means you will not be able to tell whether the numbers are in thousands or millions of dollars, percentages, days, raw numbers, etc. In addition, the data you will receive has been normalized so that all product quality and customer satisfaction measures have comparable means and standard deviations.

The performance measures are presented this way because it is important that your analysis focuses on this data only and is not influenced by any other experiences you may have. For example, say that instead of having generic labels, two of the product quality measures were labeled *% defects in production* and *warranty costs*, and one of the customer satisfaction measures was labeled *sales from repeat customers*. For someone who works in an organization in which warranty costs are important, their experience may lead them to assume that *warranty costs* are a better predictor of *sales from repeat customers* than is *% defects in production*, but that may not be the case here. Alternately, someone who works for an organization in which production-line defects are important might assume that *% defects in production* is a better predictor of *sales from repeat customers* than is *warranty costs* based on their experiences, but again that may not be the case here. The use of generic labels and normalized data helps reduce the impact of such prior experiences in this setting.

Keep in mind two things about these generic labels.

1) Whenever a measure is labeled "product quality measure #1", "customer satisfaction measure #1", etc., these labels always refer to the same measures. In other words, "product quality measure #1" does not represent *% defects* on one page or chart and *warranty costs* on another; it always represents the same measure.

2) Each product quality measure may or may not be useful for understanding changes in one, more than one, or none of the customer satisfaction measures. In other words, just because a measure is labeled "product quality measure #1" does not imply that it should be used to make judgments about "customer satisfaction measure #1" and that none of the other customer satisfaction measures should be used.

Information About Your Organization

Your organization has a total of 40 plants. All of the plants make the same products and were built to the same design, so the production scale and technology is similar in all of them. In addition, the customer segment served by each plant is similar. Because of these similarities in products, production scale, technology, and customers, the effect of product quality on customer satisfaction is **roughly the same** across plants. However, plant managers have some freedom in how much emphasis and resources they place on different product quality activities (e.g., prevention of product quality problems, appraisal of product quality level, correction of product quality problems). In other words, the manager of Plant X may place more emphasis on the prevention of product quality problems than on the appraisal of product quality level, while the manager of Plant Y may do the opposite.

This Page for Indicator Structure Condition Only

Your Task

To see if the new performance measurement system helps you to learn about the relationship between product quality and customer satisfaction, management has randomly selected 20 of the 40 plants in your organization and is providing you with measures of product quality and customer satisfaction for those 20 plants. The data are from periods in which there were no significant external shocks (e.g., foreign currency crises, strikes, etc.), unusual internal events, or seasonal variations that would alter or mask the effects of product quality on customer satisfaction.

See what you can learn from this data about the relation between product quality and customer satisfaction. Examine it at your own pace. When you believe that you have learned all that you can about the relationship between product quality and customer satisfaction, go to the next envelope.

	Product Quality Measures		Customer Satisfaction Measures	
Plant No.	Measure #1	Measure #2	Measure #1	Measure #2
2	##.##	##.##	##.##	##.##
4	##.##	##.##	##.##	##.##
7	##.##	##.##	##.##	##.##
8	##.##	##.##	##.##	##.##
10	##.##	##.##	##.##	##.##
11	##.##	##.##	##.##	##.##
12	##.##	##.##	##.##	##.##
16	##.##	##.##	##.##	##.##
19	##.##	##.##	##.##	##.##
20	##.##	##.##	##.##	##.##
21	##.##	##.##	##.##	##.##
22	##.##	##.##	##.##	##.##
24	##.##	##.##	##.##	##.##
27	##.##	##.##	##.##	##.##
29	##.##	##.##	##.##	##.##
30	##.##	##.##	##.##	##.##
33	##.##	##.##	##.##	##.##
36	##.##	##.##	##.##	##.##
39	##.##	##.##	##.##	##.##
40	##.##	##.##	##.##	##.##

DO NOT RETURN THESE MATERIALS TO THEIR ENVELOPE YET. YOU MAY WANT TO REFER TO THEM WHILE YOU WORK ON THE MATERIALS IN THE NEXT ENVELOPE.

This Page for Construct Structure Condition Only

Your Task

To see if the new performance measurement system helps you to learn about the relationship between product quality and customer satisfaction, management has randomly selected 20 of the 40 plants in your organization and is providing you with measures of product quality and customer satisfaction for those 20 plants. The data are from periods in which there were no significant external shocks (e.g., foreign currency crises, strikes, etc.), unusual internal events, or seasonal variations that would alter or mask the effects of product quality on customer satisfaction.

On the next page you will find the set of product quality measures for the selected 20 plants. Any of these measures alone is an imperfect measure of the true level of product quality, but taken together they may help you judge what that level is. For each set of measures, management wants you to estimate what the level of product quality is for that plant. You should use a 0-100 scale, where 0 = lowest possible level of product quality and 100 = highest possible level of product quality. Your judgments of the levels of product quality do not have to be in whole numbers (i.e., your judgment could be 57.346).

	Product Quality Measures		Your Estimate of the Plant's Product Quality, Using a 0-100 Scale (0=lowest, 100=highest)
Plant No.	Measure #1	Measure #2	
2	##.##	##.##	
4	##.##	##.##	
7	##.##	##.##	
8	##.##	##.##	
10	##.##	##.##	
11	##.##	##.##	
12	##.##	##.##	
16	##.##	##.##	
19	##.##	##.##	
20	##.##	##.##	
21	##.##	##.##	
22	##.##	##.##	
24	##.##	##.##	
27	##.##	##.##	
29	##.##	##.##	
30	##.##	##.##	
33	##.##	##.##	
36	##.##	##.##	
39	##.##	##.##	
40	##.##	##.##	

How accurate do you believe your estimates of the levels of product quality are?

1	2	3	4	5	6	7	8	9	10
Extremely Inaccurate									Extremely Accurate

This Page for Construct Structure Condition Only

Now that you have estimated the relation of the product quality measures to the level of product quality for the 20 plants, management wants you to learn the relation **between** product quality and customer satisfaction for those same 20 plants.

The table below includes **the same product quality measures and values for the same plants that were in the last table**. In addition, you will also find the customer satisfaction measures for those plants.

See what you can learn from the data about the relation between product quality and customer satisfaction. Examine this data at your own pace. When you believe that you have learned all that you can about the relationship between product quality and customer satisfaction, go to the next envelope.

	Product Quality Measures		Customer Satisfaction Measures	
Plant No.	Measure #1	Measure #2	Measure #1	Measure #2
2	##.##	##.##	##.##	##.##
4	##.##	##.##	##.##	##.##
7	##.##	##.##	##.##	##.##
8	##.##	##.##	##.##	##.##
10	##.##	##.##	##.##	##.##
11	##.##	##.##	##.##	##.##
12	##.##	##.##	##.##	##.##
16	##.##	##.##	##.##	##.##
19	##.##	##.##	##.##	##.##
20	##.##	##.##	##.##	##.##
21	##.##	##.##	##.##	##.##
22	##.##	##.##	##.##	##.##
24	##.##	##.##	##.##	##.##
27	##.##	##.##	##.##	##.##
29	##.##	##.##	##.##	##.##
30	##.##	##.##	##.##	##.##
33	##.##	##.##	##.##	##.##
36	##.##	##.##	##.##	##.##
39	##.##	##.##	##.##	##.##
40	##.##	##.##	##.##	##.##

DO NOT RETURN THESE MATERIALS TO THEIR ENVELOPE YET. YOU MAY WANT TO REFER TO THEM WHILE YOU WORK ON THE MATERIALS IN THE NEXT ENVELOPE.

Envelope 3
Judgment Materials and Data

Data on Judgment Table Differs by Experimental Condition in Terms of:

1) **Number of Measures of Product Quality (2 or 5)**
2) **Multicollinearity in Product Quality Measures (Low or High)**

Data on Table for Self-Report of Weights Differs by Experimental Condition in Terms of Number of Measures of Product Quality (2 or 5)

The table below provides data on the *same* product quality measures as those you have already studied; "product quality measure #1" on this table is the same measure as "product quality measure #1" on the last table, etc. However, this data is for the *other* 20 plants in your firm. These plants are comparable to those that were listed in the prior table; they were built to the same design so their production scale and technology is similar, and the primary customers served by the plants are similar.

To see how well the new performance measurement system has helped you to learn the relation between product quality and customer satisfaction, management is asking you to make estimations using this data. In the blank column, enter your best estimations of Customer Satisfaction Measures #1 and #2 for each of these twenty plants.

	Product Quality Measures		Estimate of Customer Satisfaction Measures	
Plant No.	Measure #1	Measure #2	Measure #1	Measure #2
1	##.##	##.##		
3	##.##	##.##		
5	##.##	##.##		
6	##.##	##.##		
9	##.##	##.##		
13	##.##	##.##		
14	##.##	##.##		
15	##.##	##.##		
17	##.##	##.##		
18	##.##	##.##		
23	##.##	##.##		
25	##.##	##.##		
26	##.##	##.##		
28	##.##	##.##		
31	##.##	##.##		
32	##.##	##.##		
34	##.##	##.##		
35	##.##	##.##		
37	##.##	##.##		
38	##.##	##.##		

Please allocate 100 points across the product quality measures, based on their relative importance to your estimations of each customer satisfaction measure.

For example:

* if only Product Quality Measure #1 was important to your estimations of Customer Satisfaction Measure #1, then you should enter 100 under "Product Quality Measure #1" and zero in the remaining blank cells
* if all product quality measures were equally important to your estimations of Customer Satisfaction Measure #2, then you should allocate the 100 points evenly across all the measures.

Importance Of Product Quality Measures For Estimation Of CUSTOMER SATISFACTION MEASURE #1		
Total Points to Allocate	Product Quality Measure #1	Product Quality Measure #2
100		

Importance Of Product Quality Measures For Estimation Of CUSTOMER SATISFACTION MEASURE #2		
Total Points to Allocate	Product Quality Measure #1	Product Quality Measure #2
100		

BEFORE PROCEEDING TO THE NEXT PART OF THE TASK, RETURN ALL MATERIALS THAT YOU NOW HAVE OUT TO ONE OF THE TWO EMPTY ENVELOPES (the other envelope will remain empty).

Envelope 4
Post-Experiment Questionnaire

Materials Are the Same Across All Experimental Conditions

1) How accurate do you believe your customer satisfaction estimates were (i.e., how close do you believe your estimates were to the best possible estimates)?

1	2	3	4	5	6	7	8	9	10
Extremely Inaccurate									Extremely Accurate

2) How complex is this performance measurement system (i.e., were there a lot of interrelationships between and among the measures that you needed to consider when making your customer satisfaction estimates)?

1	2	3	4	5	6	7	8	9	10
Not At All Complex									Extremely Complex

3) How difficult was it to make the estimations of customer satisfaction (i.e., was this task hard for you to do)?

1	2	3	4	5	6	7	8	9	10
Extremely Easy									Extremely Difficult

4) How familiar are you with making judgments and estimations like those you made for customer satisfaction (i.e., have you made similar judgments or estimates in the past, either for work or school)?

1	2	3	4	5	6	7	8	9	10
Extremely Unfamiliar									Extremely Familiar

5) I based my estimates of customer satisfaction on **only some** of the product quality measures that were provided.

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

a) Please indicate the number of product quality measures you used to make your estimates. _____

b) I used different product quality measures for different customer satisfaction estimates (e.g., I might have used product quality measure #1 for some estimates and product quality measure #2 for others).

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

The questions on the next three pages ask about the how you made your estimates of customer satisfaction.

1) Using data for the first twenty plants, I **estimated a weight** to be placed on each product quality measure I wanted to use, and then **combined** the weights and the product quality measures to make my estimates of customer satisfaction.

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

☐ Check this box if 1) does not describe how you made your estimates and continue to 2) on the next page.

If 1) does describe how you made your estimates, answer a) and b) that follow.

a) How did you **estimate the weights** for the product quality measures you used?

i) One way to estimate weights is to use **equal weights** for the product quality measures, based on the number of product quality measures (i.e., if there were two measures, the weight to be placed on each would be 1/2; if there were three measures, the weight to be placed on each would be 1/3, etc.).

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

ii) Another way to estimate the weights is to choose one pair of the twenty plants for which all the product quality and customer satisfaction measures were provided, and compute how much customer satisfaction changed for every one-unit change in product quality. In other words, the measures for two selected plants would be used in the formula:

$$\frac{\text{(customer satisfaction measure at Plant A – customer satisfaction measure at Plant B)}}{\text{(product quality measure at Plant A – product quality measure at Plant B)}}$$

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

iii) I did not use either method above to estimate weights, but instead I used this method: _____

_____

_____

_____

b) How did you **combine** the weights and the values of the product quality measures to make your estimates of customer satisfaction?

i) For each product quality measure, I multiplied its weight by the value for the measure, and then I **added** those results for all the measures together (i.e., I used a weighted average combination).

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

ii) I did not use a weighted average combination, but instead I did this:

_____

_____

_____

_____

2) I developed a mathematical formula to combine the product quality measures into an estimate of customer satisfaction (e.g., customer satisfaction #1 = (product quality #1 / 3) + (product quality #2 / 2) + ...).

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

Please write the mathematical formula that you used:

3) I matched a given set of product quality measures from the **second** 20 plants to the table of product quality measures for the **first** 20 plants and based my estimate of customer satisfaction on the best match I found.

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

4) I calculated the mean of customer satisfaction for the **first** 20 plants and used that as my estimate of customer satisfaction for the **second** 20 plants.

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

5) I used the same value for each customer satisfaction estimate.

1	2	3	4	5	6	7	8	9	10
Not At All Descriptive of How I Made My Estimates									Exactly Describes How I Made My Estimates

Please describe how you determined the value you used:

_____

_____

_____

_____

6) If none of the statements above describe how you made your estimates of customer satisfaction, or if you would like to describe what you did in more detail, please do so here.

_____

_____

_____

_____

The questions on this page ask for more general information about how you made your customer satisfaction estimates.

1) I could not decide on one single way/approach to make my estimates of customer satisfaction, so I used a combination of different approaches.

1	2	3	4	5	6	7	8	9	10
Strongly Disagree									Strongly Agree

2) I know that actual customer satisfaction outcomes cannot be perfectly predicted by a mathematical model, so I based my customer satisfaction estimates on a mathematical model I developed but made plant-by-plant adjustments to those estimates.

1	2	3	4	5	6	7	8	9	10
Strongly Disagree									Strongly Agree

3) How did the following features of the product quality and customer satisfaction measures affect your attitude about making the customer satisfaction estimates?

	I Really Did Not Like This									I Really Did Like This
a) generic names for performance measures?	1	2	3	4	5	6	7	8	9	10
b) comparable means and standard deviations across all performance measures?	1	2	3	4	5	6	7	8	9	10

4) Do you think it would have been easier or more difficult to make your customer satisfaction estimates if the product quality and customer satisfaction data had:

	Extremely Easier									Extremely More Difficult
a) more specific names for performance measures?	1	2	3	4	5	6	7	8	9	10
b) different means and standard deviations across all measures?	1	2	3	4	5	6	7	8	9	10

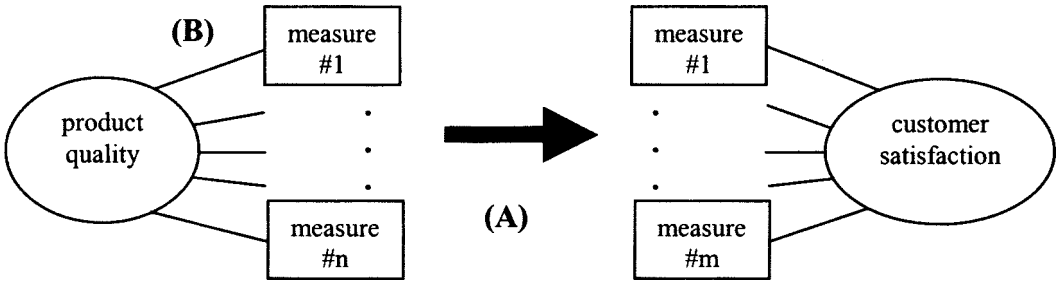
5) Would you have changed the way you made your estimates of customer satisfaction if the product quality and customer satisfaction measures had more specific names and/or different means and standard deviations? yes no

If yes, what would you have done differently? _____

_____

The questions on the next two pages ask your beliefs about and experiences with the performance measurement system used to make your customer satisfaction estimates.

Assume that an organization is using the performance measurement system illustrated below, which is reproduced from the materials you used.



		Extremely Unfamiliar									Extremely Familiar
1)	How familiar are you with performance measurement systems like the one pictured?	1	2	3	4	5	6	7	8	9	10

YOUR BELIEFS ABOUT THE RELATIONS AT (A) IN THE PICTURE:

		Strongly Disagree									Strongly Agree
2)	Because management assumes that product quality affects customer satisfaction, **each** measure of product quality is significantly correlated with **each** measure of customer satisfaction.	1	2	3	4	5	6	7	8	9	10
3)	Because management assumes that product quality affects customer satisfaction, **each** measure of product quality is significantly correlated with **at least one but not all** of the customer satisfaction measures.	1	2	3	4	5	6	7	8	9	10
4)	When I was making my estimates of customer satisfaction, I focused my attention on estimating relationships **between the measures of product quality and the measures of customer satisfaction** (i.e., the relations at (A) in the picture).	1	2	3	4	5	6	7	8	9	10

YOUR BELIEFS ABOUT THE RELATIONS AT (B) IN THE PICTURE:

		Strongly Disagree									Strongly Agree
5)	If an organization has **more** measures of product quality as opposed to fewer measures, this indicates that product quality is **more important to the organization's long-term goals.**	1	2	3	4	5	6	7	8	9	10
6)	If an organization has **more** measures of product quality as opposed to fewer measures, this indicates that **each quality measure individually is less accurate.**	1	2	3	4	5	6	7	8	9	10
7)	When I was making my estimates of customer satisfaction, I focused my attention on estimating relationships **between the measures of product quality** (i.e., the relations at (B) in the picture).	1	2	3	4	5	6	7	8	9	10
8)	The relationships **between** the product quality measures themselves (i.e., the relations at (B) in the picture) influence the relationships between the product quality and the customer satisfaction measures (i.e., the relations at (A)).	1	2	3	4	5	6	7	8	9	10
9)	Although the relationships **between** the product quality measures themselves (i.e., the relations at (B)) have an impact on the relationships between the product quality and the customer satisfaction measures (i.e., the relations at (A)), I do not know how to incorporate this into my estimates.	1	2	3	4	5	6	7	8	9	10
10)	I thought that some or all of the product quality measures I used to make my estimates **were highly correlated with each other**.	1	2	3	4	5	6	7	8	9	10

Answer EITHER a) or b) below, depending on your answer to 10). DO NOT answer BOTH a) and b).

a) Because **I DID NOT** think the product quality measures were highly correlated with each other:

I thought some or all were measuring something other than product quality.	1	2	3	4	5	6	7	8	9	10
I thought some or all were measuring different dimensions of product quality.	1	2	3	4	5	6	7	8	9	10

b) Because **I DID** think that some or all of the product quality measures were highly correlated with each other:

I determined which of the product quality measures explained the most change in customer satisfaction, and I used only those when making my estimates.	1	2	3	4	5	6	7	8	9	10
I had a difficult time determining which of the product quality measures explained the most change in customer satisfaction, so I used different measures for different estimates.	1	2	3	4	5	6	7	8	9	10

The questions on the next three pages concern your knowledge of statistics.

1) Suppose a reliable statistical analysis for an industry shows a high significant positive correlation, *r*, between variables X and Y. Which of the following statements can we then conclude is **TRUE**?

   a) On average, organizations in this industry that have higher levels of X have higher levels of Y, and organizations in this industry that have lower levels of X have lower levels of Y.

   b) On average, organizations in this industry have high X values and high Y values.

   c) It is impossible for an organization in this industry to have a high X value and a low Y value.

   d) X has no predictable association with Y in this industry.

   e) All of the above are true.

   f) None of the above are true.

2) Which of the following statements about the correlation coefficient, *r*, is **TRUE**?

   a) The correlation coefficient, *r*, measures the degree of linear or nonlinear relationship between two variables.

   b) An *r* value of 0.02 indicates a very high level of correlation between two variables.

   c) The correlation coefficient, *r*, for X and Y measures the strength and direction of the relationship between the variables.

   d) The correlation coefficient, *r*, for variables X and Y is always the same as the slope coefficient *b* for X when Y is regressed on X and several other independent variables in a multiple regression.

   e) All of the above are true.

   f) None of the above are true.

3) As the relationship between two variables, X and Y, decreases from a correlation, *r*, of 1.0, what happens to the X-Y points on a scatter diagram of the two variables?

   a) They become more scattered.
   b) The slope changes.
   c) The intercept changes.
   d) All of the above.
   e) None of the above.

4) If two variables have a correlation coefficient, *r*, of 0.30, what percentage of the change in one variable is accounted for by changes in the other variable?

   a) 60%
   b) 30%
   c) 15%
   d) 9%
   e) None of the above.

5) Suppose that materials handling costs (in dollars) for a manufacturing plant can be well represented by a regression model of the form:

$y = a + b_1x_1 + b_2x_2$ where y = material handling costs
x_1 = number of material moves
x_2 = number of pounds of material moved

Estimation of this model with recent data from the plant provides the following coefficient estimates (all significant at $p<.05$). These are unstandardized coefficients (i.e., they are stated in terms of dollars, not standard deviations).

$a = \$40{,}000$ $b_1 = \$2.00$ $b_2 = \$0.10$

What would you expect material handling costs to be in a quarter when 500 moves were made and a total of 10,000 pounds of material was moved?

a) $40,002
b) $42,000
c) $40,000
d) $2,000
e) None of the above.

6) In regression analysis, observed errors, which represent information from the data which is not explained by the model, are called:

a) marginal values
b) residuals
c) mean square errors
d) standard errors
e) none of the above.

7) Which of the following statements about multiple regression is TRUE?

a) When doing individual *t*-tests on each of the independent variables, X, in a multiple regression model, each test is independent of each other test.

b) Using multiple regression to regress five independent X variables to predict Y will give the same result as five separate regressions of Y on each independent X variable.

c) Adding an independent variable to the model can never reduce the **unadjusted** R^2.

d) Adding an independent variable to the model can never reduce the **adjusted** R^2.

e) None of the above is true.

8) When the null hypothesis H_0: $B_1 = B_2 = B_3 = 0$ is rejected, the interpretation should be:

a) there is no linear relationship between Y and any of the three independent X variables.
b) there is a relationship between Y and at least one of the three independent X variables.
c) all three independent X variables have a slope not significantly different than zero.
d) all three independent X variables have equal slopes.
e) none of the above.

9) Multicollinearity can be described as:

a) a regression model with more than one independent variable.
b) a regression model with circular relations between the independent variables.
c) a regression model with correlations between the independent variables.
d) a regression model with exponential variables.
e) none of the above.

10) In regression analysis, all of the following are possible effects of multicollinearity EXCEPT:

a) estimated regression coefficients for the independent variables in the model remain the same even when some of the independent variables are removed from the model.

b) the signs of the estimated regression coefficients for the independent variables may be the opposite of what is expected.

c) a significant *F* ratio for the regression model may result even though the *t* ratios for each independent variable are not significant.

d) the variances (standard errors) of the regression coefficients estimates for the independent variables can be larger than expected.

e) none of the above.

11) You manufacture two kinds of candy, and you want to regress monthly operating costs on volumes of the ingredients used in the same month. Your candy recipes are as follows:

for one lb. of Candy 1: 1/4 lb. dark chocolate + 1/3 lb. nuts

for one lb. of Candy 2: either 1/2 lb. milk chocolate + 1/4 lb. cocoa
or 1/3 lb. milk chocolate + 1/2 lb. cocoa
(machine operators vary the recipe according to availability of ingredients)

Which of the following regression models would NOT have multicollinearity problems? Note that you are **not** being asked to select the best model.

a) operating costs = $b_0 + b_1$(lbs. dark chocolate) + b_2(lbs. nuts) + e

b) operating costs = $b_0 + b_1$(lbs. dark chocolate) + b_2(lbs. nuts) + b_3(lbs. of cocoa) + e

c) operating costs =
$b_0 + b_1$(lbs. dark chocolate) + b_2(lbs. nuts) + b_3(lbs. milk chocolate) + b_4(lbs. cocoa) + e

d) operating costs = $b_0 + b_1$(lbs. dark chocolate) + b_2(lbs. milk chocolate) + b_3(lbs. cocoa) + e

e) none of the above.

These final questions ask about your school and work background.

1) What is your area(s) of concentration in the MBA program? _____

2) Before you entered the MBA program, what was your last job title?_____

3) In what industry(ies) did you work before you entered the MBA program (e.g., automotive, brokerage, retail)? _____

4) In what managerial functional area(s) did you work before you entered the MBA program (e.g., accounting, marketing, sales, finance, production)? _____

5) How many months of full-time equivalent managerial work experience did you have before you started the MBA program? _____ months

6) Please complete the following table about coursework you have completed before you began the MBA program.

	number of semester-length 3-credit-hour courses (or their equivalent)	average GPA in those courses (max = 4.0)
statistics		
mathematics		
accounting		
finance		
supply chain management		
quality management		
operations management		

7) How familiar are you with the following statistical analysis tools?

		Extremely Unfamiliar									Extremely Familiar
a)	Factor analysis	1	2	3	4	5	6	7	8	9	10
b)	Structural equation modeling (i.e., SEM, LISREL, causal modeling)	1	2	3	4	5	6	7	8	9	10

BIBLIOGRAPHY

Armelius, B. and K. Armelius. 1974. The use of redundancy in multiple-cue judgments: Data from a suppressor-variable task. *American Journal of Psychology* 87: 385-392.

Armelius, K. and B. Armelius. 1975. Note on detection of cue intercorrelation in multiple-cue probability learning. *Scandinavian Journal of Psychology* 16: 37-41.

Ashton, R. 1982. *Human Information Processing in Accounting*. Sarasota, FL: American Accounting Association.

Ashton, R. 1990. Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research* 28 Supplement: 148-180.

Balkcom, J., C. Ittner, and D. Larcker. 1997. Strategic performance measurement: Lessons learned and future directions. *Journal of Strategic Performance Measurement* 1: 22-32.

Banker, R., G. Potter, and D. Srinavasan. 2000. An empirical investigation of an incentive plan that includes nonfinancial measures. *The Accounting Review* 75: 65-92.

Barron, F. 1988. Limits and extensions of equal weights in additive multiattribute models. *Acta Psychologica* 68: 141-152.

Bettman, J., E. Johnson and J. Payne. 1990. A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes* 45: 111-139.

Bloomfield, R., R. Libby and M. Nelson. 1998a. Over-reliance on previous periods' earnings can cause post-earnings-announcement drift and over-reactions to extreme performance. Working paper.

Bloomfield, R., R. Libby and M. Nelson. 1998b. Underreactions and overreactions: The influence of information reliability and portfolio formation rules. Working paper.

Bonner, S. 1994. A model of the effects of audit task complexity. *Accounting, Organizations and Society* 19: 213-234.

Brehmer, B. 1971. Subjects' ability to use functional rules. *Psychonomic Science* 24: 259-260.

Brehmer, B. 1973a. Note on the relation between single-cue probability learning and multiple-cue probability learning. *Organizational Behavior and Human Performance* 9: 246-252.

Brehmer, B. 1973b. Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior and Human Performance* 9: 377-395.

Brehmer, B. 1974a. Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance* 11: 1-27.

Brehmer, B. 1974b. The effect of cue intercorrelation on interpersonal learning of probabilistic inference tasks. *Organizational Behavior and Human Performance* 12: 397-412.

Brehmer, B. 1987. Note on subjects' hypotheses in multiple-cue probability learning. *Organizational Behavior and Human Decision Processes* 40: 323-329.

Brehmer, B., J. Kuylenstierna, and J. Liljergren. 1974. Effects of function form and cue validity on the subjects' hypotheses in probabilistic inference tasks. *Organizational Behavior and Human Performance* 11: 338-354.

Broniarczyk, S. and J. Alba. 1994. Theory versus data in prediction and correlation tasks. *Organizational Behavior and Human Decision Processes* 57: 117-139.

Casey, C. 1980. Variation in accounting information load: The effect on loan officers' predictions of bankruptcy. *The Accounting Review* 55: 36-49.

Chewning, E. and A. Harrell. 1990. The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society* 15: 527-542.

Cohen, J. and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawes, R. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34: 571-582.

Dawes, R. and B. Corrigan. 1974. Linear models in decision making. *Psychological Bulletin* 81: 95-106.

Ernst & Young, LLP. 2002. Documentation and personal communication with senior manager regarding personnel performance measurement system.

Getzels, J. 1982. The problem of the problem. In R. Hogarth (Ed.), *Question Framing and Response Consistency*. San Francisco, CA: Jossey-Bass, Inc.

Goodwin, P. and G. Wright. 1993. Improving judgmental times series forecasting: A review of the guidance provided by research. *International Journal of Forecasting* 9: 147-161.

Goodwin, P. and G. Wright. 1994. Heuristics, biases and improvement strategies in judgmental time series forecasting. *Omega* 22: 553-568.

Greene, W. 2000. *Econometric Analysis, Fourth Edition.* Upper Saddle River, NJ: Prentice-Hall.

Gujarati, D. 1988. *Basic Econometrics, Second Edition*. New York, NY: McGraw-Hill.

Hertenstein, J. and M. Platt. 2000. Performance measures and management control in new product development. *Accounting Horizons* 14: 303-323.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74-91.

Huber, V. 1985. Effects of task difficulty, goal setting, and strategy on performance of a heuristic task. *Journal of Applied Psychology* 70: 492-504.

Hutchinson, J. and J. Alba. 1997. Heuristics and biases in the "eyeballing" of data: The effects of context on intuitive correlation assessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23: 591-621.

Iselin, E. 1988. The effects of information load and information diversity on decision quality in a structured decision task. *Accounting, Organizations and Society* 13: 147-164.

Ittner, C. and D. Larcker. 1998. Innovations in Performance Measurement: Trends and Research Implications. *Journal of Management Accounting Research* 10: 205-238.

Ittner, C., D. Larcker and M. Meyer. 2002. Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. Working paper.

Jiambalvo, J. and W. Waller. 1984. Decomposition and assessments of audit risk. *Auditing: A Journal of Practice and Theory* 3: 80-88.

Kaplan, R. and D. Norton. 1992. The balanced scorecard - measures that drive performance. *Harvard Business Review* 70: 71-79.

Kaplan, R. and D. Norton. 1993. Putting the balanced scorecard to work. *Harvard Business Review* 71: 134-147.

Kaplan, R. and D. Norton. 1996a. Linking the balanced scorecard to strategy. *California Management Review* 39: 53-79.

Kaplan, R. and D. Norton. 1996b. *The Balanced Scorecard: Translating Strategy Into Action*. Boston, MA: Harvard Business School Press.

Kaplan, R. and D. Norton. 1996c. Using the balanced scorecard as a strategic management system. *Harvard Business Review* 74: 75-85.

Kaplan, R. and D. Norton. 2000. Having trouble with your strategy? Then map it. *Harvard Business Review* 78: 167-176.

Kaplan, R. and D. Norton. 2001. *The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment.* Boston, MA: Harvard Business School Press.

Kaplan, R. and N. Tempest. 1999. *Wells Fargo Online Financial Services (A).* Boston, MA: Harvard Business School Press, case 9-198-146.

Kennedy, P. 2001. *A Guide to Econometrics, Fourth Edition.* Cambridge, MA: The MIT Press.

Klayman, J. 1988. On the how and why (not) of learning from outcomes. In B. Brehmer and C. Joyce (Eds.), *Human Judgment: The SJT View.* Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Krumwiede, K., T. Eaton, and M. Swain. 2000. The effects of strategic linkages, evaluation focus, and financial outcomes on performance evaluations in a balanced scorecard framework. Working paper.

Lambert. R. 1998. Customer satisfaction and future financial performance - discussion of "Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction." *Journal of Accounting Research* 36: 37-46.

Laughlin, J. 1978. Comment on "Estimating coefficients in linear models: It don't make no nevermind." *Psychological Bulletin* 85: 247-253.

Lee, J. and J. Yates. 1992. How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin* 112: 363-377.

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

Lindell, M. and T. Stewart. 1974. The effects of redundancy in multiple-cue probability learning. *American Journal of Psychology* 87: 393-398.

Lipe, M. and S. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review* 75: 283-298.

Lipe, M. and S. Salterio. 2002. A note on the judgmental effects of the balanced scorecard's information organization. *Accounting, Organizations & Society* 27: 531-540.

Luft, J. and M. Shields. 2001. Why does fixation persist? Experimental evidence on the judgment performance effects of expensing intangibles. *The Accounting Review* 76: 561-587.

Maines, L. 1990. The effect of forecast redundancy on judgments of a consensus forecast's expected accuracy. *Journal of Accounting Research* 28 Supplement: 29-47.

Maines, L. 1996. An experimental examination of subjective forecast combination. *International Journal of Forecasting* 12: 223-233.

Merlo, A. and A. Schotter. 2001. Learning by not doing: An experimental investigation of observational learning. Penn Institute for Economic Research Working Paper Number 01-040.

Messier, W. 1995. Research in and development of audit decision aids. In R. Ashton and A. Ashton (Eds.), *Judgment and Decision-Making Research in Accounting and Auditing*. Cambridge, United Kingdom: Cambridge University Press.

Miller, P. 1971. Do labels mislead? A multiple cue study, within the framework of Brunswik's probabilistic functionalism. *Organizational Behavior and Human Performance* 6: 480-500.

Naylor, J. and R. Clark. 1968. Intuitive inference strategies in interval learning tasks as a function of validity magnitude and sign. *Organizational Behavior and Human Performance* 3: 378-399.

Naylor, J. and E. Schenck. 1968. The influence of cue redundancy upon the human inference process for tasks of varying degrees of predictability. *Organizational Behavior and Human Performance* 3: 47-61.

Neter, J., M. Kutner, C. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Statistical Models, Fourth Edition*. Chicago, IL: Irwin.

Nisbett, R., H. Zukier and R. Lemley. 1981. The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology* 13: 248-277.

Payne, J. 1976. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance* 16: 366-387.

Payne, J., J. Bettman, and E. Johnson. 1988. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 534-552.

Payne, J., J. Bettman, and E. Johnson. 1990. The adaptive decision maker: Effort and accuracy in choice. In R. Hogarth (Ed.), *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. Chicago, IL: The University of Chicago Press.

Payne, J., J. Bettman, and E. Johnson. 1992. Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology* 43: 87-131.

Peterson, C., K. Hammond and D. Summers. 1965. Optimal responding in multiple-cue probability learning. *Journal of Experimental Psychology* 70: 270-276.

Pruzek, R. and B. Frederick. 1978. Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin* 85: 254-266.

Ruscio, J. 2000. The role of complex thought in clinical prediction: Social accountability and the need for cognition. *Journal of Consulting and Clinical Psychology* 68: 145-154.

Schmitt, N. and A. Dudycha. 1975. A reevaluation of the effect of cue redundancy in multiple-cue probability learning. *Journal of Experimental Psychology* 104: 307-315.

Schum, D. and A. Martin. 1982. Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review* 17: 105-151.

Shields, M. 1980. Some effects of information load on search patterns used to analyze performance reports. *Accounting, Organizations and Society* 5: 429-442.

Shields, M. 1983. Effects of information supply and demand on judgment accuracy: Evidence from corporate managers. *The Accounting Review* 58: 284-303.

Simon, H. 1978. Information-processing theory of human problem solving. In W. Estes (Ed.), *Handbook of Learning and Cognitive Processes, Volume 5: Human Information Processing*. Hillsdale, NJ: Erlbaum Associates.

Simons, R. and A. Davila. 1998. How high is your return on management? *Harvard Business Review* 76: 70-80.

Sjoblom, L. 1998. Financial information and quality management: Is there a role for accountants? *Accounting Horizons* 12: 363-373.

Slovic, P., D. Griffin, and A. Tversky. 1990. Compatibility effects in judgment and choice. In R. Hogarth (Ed.), *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. Chicago, IL: The University of Chicago Press.

Sprinkle, G. 2002. Perspectives on experimental research in managerial accounting. *Accounting, Organizations and Society* Forthcoming.

Stivers, B., T. Covin, N. Hall, and S. Smalt. 1998. How nonfinancial performance measures are used. *Management Accounting* 79: 44-49.

Trabasso, T. 1982. The importance of context in understanding discourse. In R. Hogarth (Ed.), *Question Framing and Response Consistency*. San Francisco, CA: Jossey-Bass, Inc.

Tuttle, B. and F. Burton. 1999. The effects of a modest incentive on information overload in an investment analysis task. *Accounting, Organizations and Society* 24: 673-687.

Tversky, A., S. Sattath, and P. Slovic. 1988. Contingent weighting in judgment and choice. *Psychological Review* 95: 371-384.

Ullrich, M. and B. Tuttle. 2000. The effects of the balanced scorecard's information reporting system and economic incentives on effort allocation among multiple goals. Working paper.

Von Eye, A. and C. Schuster. 1998. *Regression Analysis for Social Sciences*. San Diego, CA: Academic Press.

Wainer, H. 1976. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin* 83: 213-271.

Well, A., S. Boyce, R. Morris, M. Shinjo, and J. Chumbley. 1988. Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory & Cognition* 16: 271-280.

Wood, R. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37: 60-82.