ESTIMATING THE IMPACT OF THIRD-PARTY EVALUATOR TRAINING AND CHARACTERISTICS ON THE SCORING OF WRITTEN ORGANIZATIONAL SELF-ASSESSMENTS

by

Garry D. Coleman

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial and Systems Engineering

APPROVED:

C. Patrick Koelling, Chairman

Lawrence H Cross

Lawrence Cross

E Scott Geller

E. Scott Geller

D. Scott Sink

D. Scott Sink

Konstantinos P. Triantis

Kostas Triantis

July, 1996

Blacksburg, Virginia

Key Words: Organizational Assessment, Evaluator Training, Baldrige Award

ESTIMATING THE IMPACT OF THIRD-PARTY EVALUATOR TRAINING AND CHARACTERISTICS ON THE SCORING OF WRITTEN ORGANIZATIONAL SELF-ASSESSMENTS

by

Garry D. Coleman
C. Patrick Koelling, Chairman
Industrial and Systems Engineering

(ABSTRACT)

This study examined the process of third-party scoring of organizational self-assessments. An experiment was conducted to illustrate the magnitude of score consistency and accuracy among evaluators, estimate the impact of frame-of-reference (FOR) training on score consistency and accuracy, and explore the relationship between evaluator characteristics and score accuracy. The organizational self-assessment used was the 1995 Malcolm Baldrige National Quality Award Colony Fasteners Case Study. The subjects were 81 graduate students enrolled in two televised graduate engineering courses with considerable quality management content.

Subjects were randomly assigned to groups and randomly assigned to four of the seven categories of the Baldrige Award. Each subject evaluated the case study against two categories prior to the treatment. Subjects in the control group evaluated two additional categories and then a two and one-half hour FOR training intervention was provided to all subjects. Next, subjects in the treatment group evaluated their two additional categories. Finally, a questionnaire was administered regarding evaluator characteristics related to previous experience and education.

Accuracy was assessed by comparing subjects' scores to experts' scores and calculating indices (elevation and dimensional accuracy) for each subject's scores on each category. Prior to training, no statistical differences were found between groups, but a leniency effect was observed for all subjects. Category 6.0, Business Results, and Category 7.0, Customer Focus and Satisfaction, had statistically smaller score variances than the other five categories.

After training, group x time ANOVAs found evidence of an interaction. Examination of simple effects found significant differences between the group mean scores for all three items from Category 6.0 and two of the four items from Category 5.0. Significant simple time effects were found for all three items from Category 6.0 for the treatment group. No meaningful differences were found between group score variances. A significant difference in category score variance was seen across categories for the untrained group. Training improved elevation accuracy, but no evidence was seen of effects on DA.

Exploratory regression produced a prediction equation for DA with an adjusted R-square of 0.538. Predictors included work experience, QA/QC experience, employer's industry and employer's size.

Contents

Appendices

List of Figures

List of Tables

I. Introduction

The Research Problem

This research project addresses the issues of evaluator consistency and accuracy for third-party scoring of organizational self-assessments. Organizational self-assessments are a measurement and evaluation tool frequently used for diagnosis, not unlike an annual physical examination. The increasing use of organizational self-assessments leads to questions regarding their "psychometric[1]" properties. Before further describing the research problem, the context will be described by discussing why self-assessments are used, defining self-assessment and related terms, illustrating how self-assessment fits into a larger management system, and how third-party review and scoring is used to enhance the value of the self-assessment.

Improving quality and productivity is critical for the long term survival of most organizations, both in the private and public sectors. Continually striving to improve quality and productivity[2] is increasingly seen as a regular part of managing an organization. Not only is continual improvement important due to its perceived effects on financial performance, but also its impact on the effectiveness of product and service delivery. David Garvin (1988) has done an excellent job of tracing the evolution of quality improvement from simple inspection to strategic quality management. As quality improvement has been viewed as more important and strategic, the number and

[1] "The standards used to judge the quality of measures are often labeled *psychometric properties* when they refer specifically to psychological tests. On the other hand, these standards are relevant for a wide variety of measures, many of which may not appear to be very psychological in nature" (McCormick & Ilgen, 1985, p. 114).

[2] The continual improvement concept has been given many labels: total quality management (TQM), strategic quality management, continuous improvement management, etc.

sophistication of quality management tools has also increased. One of these tools or approaches is the use of periodic organizational self-assessments.

What is self-assessment? Gallagher (1994, p. 93) defined self-assessment as the "systematic measurement and review of all the key activities of the organization." Self-assessment involves the conduct and evaluation of a periodic self-study[3]. Self-assessment is different from the routine measurement and evaluation of organizational performance. Whereas measurement and evaluation of performance is often continual (e.g., weekly, monthly, quarterly), self-assessment is a "snap-shot" (e.g., once a year or every few years). Whereas continual measurement and evaluation may focus on the processes and results of ongoing operations (e.g., see Kaplan and Norton's (1992) balanced scorecard concept), self-assessment is likely to examine results, operational processes, and the improvement process. It is the evaluation of the improvement process that makes self-assessment particularly applicable for strategic quality management.

Self-assessment is of limited value if it is done in isolation. That is, self-assessment must be integrated into the management of the organization. How this integration occurs is influenced by the purpose of doing the self-assessment. Conti (1994) identified two primary purposes for self-assessment: improvement-oriented self-assessment, a diagnostic process providing the basis for new strategic improvement planning; and conformity self-assessment, an inspection oriented process designed to check. Self-assessment is often done for some combination of these two purposes. The primary focus of this research involves improvement-oriented self-assessment; however, the results may be equally applicable to conformity self-assessment. A model has been developed that can be used to describe the role of either type of self-assessment (see Figure

[3] Self-study refers to the collection of data or measurement. Self-assessment implies both measurement and evaluation.

1). The model illustrates self-assessment in the context of self-regulation. "Taken broadly, *self-regulation* refers to the actions of any individual or group to monitor and control its own behavior. ... For organizational behavior, this includes steps to be taken by a business firm, college, or other organization to monitor and control its own actions" (El-Khawas, 1983, p. 58). Others' definition of self-assessment may include all the components of self-regulation shown in Figure 1. In the model (i.e., Figure 1) self-assessment is defined in the narrower sense.

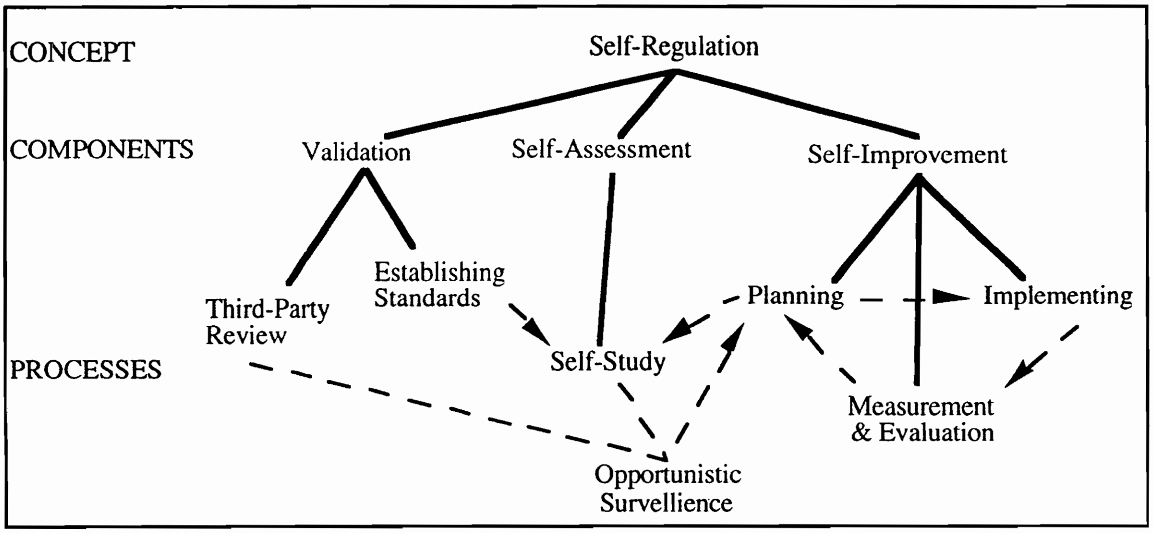


Figure 1. The processes of self-regulation.

Self-Regulation links the measurement and evaluation of self-assessment with the decisions and actions of self-improvement. The self-improvement component might be viewed as the organization's version of the Shewhart Cycle, or plan-do-study-act. Without linkage to the self-improvement component, self-assessment cannot meet Conti's improvement-oriented purpose and only meets Conti's conformity purpose in a static sense.

The model separates validation from self-assessment to emphasize its importance. Validation includes the establishment or selection of standards[4] and a third-party review process. For improvement-oriented self-assessment, it is important that the standards reflect the evolution of strategic quality management. The standards must provide guidance for identifying improvement opportunities. For conformity self-assessment, the standards must be accepted[5] by the relevant stakeholders of the organization. On the other hand, establishing standards provides an opportunity to respond proactively to issues of public concern and may reduce the need for externally imposed standards. Depending on the organization's desired outcomes, validation may or may not include a third-party[6] review process. A third-party review provides an outside perspective and may identify improvement opportunities missed by the self-study[7]. Use of a third-party demonstrates management's commitment to a serious assessment and improvement. Use of a third-party may also reinforce or validate the findings of the self-study. For conformity self-assessment, use of a third-party review demonstrates objectivity and may even be required to validate the findings of the self-study. For improvement or conformity, the third-party may be used to score the organization against the evaluation standards. Such a score may be used as a baseline for future comparisons or as an indicator of meeting some minimum standard of conformance.

[4] Standards may also be called guidelines or criteria. They provide structure for the self study, identifying areas that should be addressed. They may also include best practices or minimum acceptable performance levels. The organization may establish standards to be utilized in its self-study or it may choose to use third-party standards. In either case, there are legal, societal, professional, or institutional constraints on what are considered appropriate standards. In some cases the standards may need to be acceptable to peers and defensible to the public.

[5] In some cases the standards may be imposed, either by headquarters, regulators, or legislation.

[6] A third party is someone from outside the unit-of-analysis subject to the self-assessment that does not have a vested interest in the outcome of the review. The third party should be of sufficient status to hold the respect of the subject organization's managers and knowiedgeable of the standards being applied.

[7] Third-party review typically includes a complete review of the written self-study and may also include direct observation (i.e., a site visit) to confirm or clarify issues from the written self-study. The focus of this research is on the first review, that of the written self-study.

Conti's purposes of self-assessment are not mutually exclusive. Conti considers quality awards as conformity self-assessments, yet many of the applicants for these awards cite the improvement that results from the application process to be the primary benefit. Academic accreditation and supplier certification are also forms of conformity self-assessment, but both claim improvement as one of their purposes. Improvement-oriented self-assessment can also be used to assure stakeholders that conformity self-assessment is unnecessary and redundant. By linking self-assessment to self-improvement and validation, self-regulation may be used for both improvement and conformity purposes.

For the remainder of this document, self-assessment will be used to refer to improvement-oriented self-assessment; however, some improvement-oriented self-assessment may include conformity as a secondary purpose[8]. Self-assessment is assumed to be in the context of the self-regulation model. That is, the use of the term self-assessment implies linkage to validation and self-improvement, unless otherwise stated. This convention is being adopted to reduce confusion with the various uses found in the literature.

The increased use of self-assessment can be widely cited. Intuitively, periodic introspection followed by third-party review and a linkage to plans of action seems desirable for most any organization. The increased use of self-assessment is not without problems. As self-assessment is used as a basis for decision-making, questions regarding its effectiveness naturally arise. If an organization prepares an annual self-assessment with scoring by a third-party evaluator[9], how much of the year-to-year change in score is due to

[8] I disagree with Conti's characterization of quality awards as conformity self-assessment, particularly when referring to internal quality awards. I believe the *primary* purpose of internal quality awards is to improve the performance of the organization's subsidiaries, not to promote conformance with a common quality improvement model.

[9] Here the term evaluator will be used to refer to a third-party (outside the unit of analysis being evaluated) asked to review *and* score the organization's written self-assessment.

error and how much is due to true improvement or degradation in performance? If a team of evaluators is used, how much score variation might be expected between evaluators? Similarly, if an individual evaluator is used and chosen from a pool of evaluators, how much might the score differ depending on which evaluator is selected? How much variation in scores might be expected if the evaluator(s) prepare themselves only using the information provided by the written standards? How much might this variation be reduced by seeking additional information or training for the evaluator(s) beyond that provided in the written standards? All of these questions relate to the issue of potential error in the evaluators' scores.

Common practices used to address the issue of potential error in evaluators' scores include evaluator training and using selection criteria to create a heterogeneous pool of evaluators (Myers and Heller, 1995; Godfrey and Myers, 1994; National Institute of Standards and Technology, 1994b). Evaluator training can range from "three days of intense training" (Godfrey and Myers, 1994) provided to the AT&T Chairman's Quality Award examiners[10] or the three day course provided to Baldrige examiners (NIST, 1994b) to as little as an informal review of the criteria[11] (Ritter, 1993). No published data have been found on the effectiveness of such training.

Industrial and organizational (I/O) psychology research has found rater[12] training generally reduces rater error when assessing the performance of individuals (Stamoulis & Hauenstein, 1993; McIntyre, Smith, & Hassett, 1984; Bernardin & Walter, 1977). Generalizing the I/O psychology findings to evaluators of organizational self-assessments

[10] Examiners is a specific term used by the Malcolm Baldrige National Quality Award and the AT&T Chairman's Quality Award to refer to those who perform the role previously described as evaluator.

[11] While no training may be theoretically possible, it is unlikely an evaluator would attempt to score a self-assessment without at least familiarizing himself or herself with the criteria to be used.

[12] A rater is analogous to an evaluator, except that a rater is evaluating individuals. The term rater is frequently used in the context of performance appraisal research.

is tentative at best. The rater training studies often disagree over the measures used to represent rater error, the length of training sessions are sometimes very brief (e.g., fifteen minutes), the subjects tend to be college sophomores, the raters' task may be the performance appraisal of a brief videotape of individual behavior, and the desirability of the effect of the training differs depending on the content of the training and the measures selected to represent rater error.

A major problem with rater training research is the estimation of true scores (i.e., right answers) to which to compare the subjects' scores. Estimation procedures have included using the mean score of graduate student expert raters (McIntyre et al., 1984) and a normative approach using the mean scores from a large pilot study of undergraduate raters similar to the subjects (Stamoulis & Hauenstein, 1993). While the results of these studies may yield statistical significance, it is difficult to interpret their practical significance. Differences in scores on rating scales developed for an artificial performance appraisal task have limited meaning outside the context of the experiment. A study of the effects of evaluator training using a widely accepted scoring system and having the scores of acknowledged experts for estimates of true scores offers several improvements over similar research on rater training. A widely used scoring system allows evaluation of the practical significance of variation in pre-training scores, as well as the improvement in scores attributable to training. Using the scores of acknowledged experts as estimates of true scores supports the evaluation of the practical significance of evaluator error and the effectiveness of evaluator training.

Creating a heterogeneous pool of examiners appears sensible for a quality award, given the heterogeneous nature of the applicant organizations. Does this imply that a specific organization or industry should use a heterogeneous pool of evaluators for their self-assessments? Godfrey and Myers (1994, p. 71) describe the need to "achieve a balance" and have a "mix" of examiners for the AT&T Chairman's Quality Award. If an evaluator team approach is used, intuitively it makes sense to have evaluators with different characteristics. The assumption seems to be that the biases of the individuals attributable to their backgrounds are likely to cancel out or balance one another thus yielding mean scores closer to the true score. What if an organization is using a single evaluator? For a given organization, which characteristics are most likely to effect the magnitude of the error in the evaluator's scores? No published research has been found on this subject. If the characteristics of a group of evaluators can be correlated with the error of their scores, then relationships between evaluators' characteristics and the magnitude of error in their scores may be proposed.

This study examined evaluator consistency and evaluator accuracy as desirable surrogates for the reduction of variance and reduction of error, respectively. Consistency is the degree of agreement between evaluators (adapted from Bernardin & Walter, 1977). Consistency may also be viewed as the relative lack of variation between evaluators. Accuracy is measured as a function of the relative distance between an evaluator's scores and the true scores of ratee performance (adapted from Sulsky & Balzer, 1988). Accuracy may be viewed as the relative absence of error, where error is deviation from the true scores of ratee performance.

The research problem addressed by this study is multi-faceted. Capturing the essence of the problem in a single sentence is difficult. This study addressed many of the issues raised in the previous paragraphs by estimating the effect of third-party evaluator training and characteristics on the scoring of written organizational self-assessments. This included describing the magnitude of possible error when using minimally trained evaluators to score a written organizational self-assessment. The effect of evaluator training on the consistency and accuracy of evaluators' scores was estimated, with both statistical and practical significance examined. The effects of evaluator characteristics on the accuracy of the evaluators' scores was explored, and two models describing these relationships were proposed.

Another way of viewing the research problem is to think of the scoring of an organizational self-assessment by a third-party evaluator as a key process in the management and improvement of organizations. Little empirical data exist regarding this process, yet many decisions are made based upon its results. This research was intended to begin filling that gap and improve the basis upon which evaluators are selected and trained and evaluator scores are interpreted.

Expected Results of this Research

This research illustrates and examines the magnitude of variation and error possible when third-party evaluators score organizational self-assessments, both with and without the benefit of formal training. Training is expected to improve the accuracy of the evaluators' scores, and this will be demonstrated. The consistency between evaluators' scores on a given dimension (category or item) is illustrated. That training will improve consistency among evaluators is be demonstrated. Evaluator characteristics that appear to predict the accuracy of the evaluators' scores are identified. Where sufficient, this information is used to propose a model of evaluator characteristics believed to affect the accuracy of evaluators' scores. This model will be developed in a form suitable for testing in subsequent research.

This research produced a rich set of data and developed a number of tools for training and evaluating evaluators. A relatively brief (≤ 3 hours) evaluator training program was developed, including presentation materials and training exercises. A set of procedures that can be replicated to assess the consistency and accuracy of a pool of evaluators was demonstrated. These procedures may be used for both screening and improvement purposes. The documentation of this research, its data and other outputs is an important product in its own right. The dearth of published data and analyses implies much of the confidence in the scoring of organizational self-assessments is a matter of faith (i.e., not empirically based).

Significance of this Research

This research begins filling a gap in the body of knowledge regarding scoring written organizational self-assessments. The potential magnitude of variation and error in evaluator scores is illustrated, both with and without the benefit of formal evaluator training. For assessment processes that use the scores for decision making purposes (e.g., whether or not to recognize the organization for its performance) or as a baseline for comparison in subsequent assessments, having an estimate of potential evaluator error could be very useful for interpreting scores.

Generalizing from the sample used in this research may not always be valid; however, having an indication of evaluator error in a single controlled context is better than having no indication, or data from a poorly controlled context. In addition to examining scores, comparison of the justifications for the scores (e.g., strengths, areas for improvement) was also conducted. This information may be used to improve evaluator training and provide an indication of the evaluators' relative agreement regarding what is important.

Testing the effect of evaluator training on the consistency and accuracy of evaluator scores is a large, but logical, extrapolation of the rater research from industrial and organizational psychology. While training is expected to improve the consistency and accuracy of evaluators' scores, the statistical and practical significance of this improvement is unknown. This research represents a contribution to the evolution of industrial engineering from the micro-measurement of individuals to the measurement of organizational performance. The focus here is to increase the understanding of one of the processes used to measure and evaluate the performance of organizations.

An intermediate output of this research was a brief training program for evaluators that can be used by others. Such a program may be particularly useful for small to medium size organizations wishing to improve the effectiveness of their evaluators, but not having the resources of a Fortune 500 company. Bell and Wilson (1994) cited lack of resources for training evaluators as a major challenge for small organizations wishing to conduct regular self-assessments. The data collected were also used to identify potential improvements in the training program.

The proposed model of the relationships between evaluator characteristics and the accuracy of evaluator scores also begins filling a gap in the body of knowledge. The data used to build this model provide a basis upon which future research may be conducted. The lack of empirical evidence inhibited the development of a model of cause and effect. By using the correlational data from this research, a model can be proposed. The cause and effect relationships proposed by the model can then be tested by subsequent data collection. While cause and effect may never be truly established, this model begins the collection of evidence to support purported relationships.

Overview of this document

The document began by describing the research problem, its context and the issues that comprise the problem. The introductory chapter ended with a discussion of the expected results and significance of this research. Chapter II contains a review of the literature relevant to the research project, including self-assessment using quality awards and models, rater error research, strategic quality management, and other applications of self-assessment. Chapter III provides a detailed description of how this project was conducted. Chapter III presents the research problem and breaks it into specific research questions to be addressed. Where applicable, testable hypotheses were identified to address each question. Data required to support answering each question and testing each hypothesis were identified and the data analysis methods are briefly described. The experimental design and subjects used to produce the necessary data are also described. A detailed work plan is presented, including a work breakdown structure, descriptions of data collection procedures, and the data collection schedule. Chapter IV describes in detail the data analysis procedures. Chapter IV includes or references the raw data and many of the intermediate data products produced. Chapter IV includes the edited output from all the data analyses and some preliminary interpretation of the results. Chapter V describes the results of the research project and the supporting analyses. Chapter VI presents the conclusions of the research project. Chapter VI includes discussion of the implications of the results, lessons learned from conducting the project, and ideas for future research. The remainder of the document contains the bibliography and appendices. The bibliography lists all the works cited in this document, plus a few supporting works. The appendices include supporting materials, examples of tools used to conduct this research, raw data, and various outputs from the data analysis .

II. Review of Relevant Literature

The self-regulation model presented earlier (Figure 1) shows one view of how organizations can proactively assess and improve their performance. Postsecondary education has long practiced a collective[1] version of self-regulation through their accreditation processes (El-Khawas, 1983; Petersen, 1979), although some may argue that accreditation's purpose is more conformity than improvement (Kells, 1992a; Bender, 1983; Selden and Porter, 1977). Recent developments in accreditation may indicate a re-emphasis on improvement (Aldrige and Pape, 1994; Craft, 1992). The regulatory reform movement in the U.S. Congress (Newsweek, 1995) includes the increased use of conformity-oriented self-regulation as a way to delegate more regulatory control to the states. Organizations recognizing the competitive necessity of continual improvement (Garvin, 1988) are adopting self-assessment (in the context of self-regulation) as one of their tools. Unless organizations form collectives like postsecondary education has done, the question of what standards and procedures to use arises. One common answer is to use the standards and procedures of quality and productivity awards.

Self-assessment Using Quality Awards and Models

The use of quality awards and models as standards for organizational self-assessment is a relatively new phenomena and the research appears to be lagging practice. In 1994, the First European Forum on Quality Self-Assessment was held. The theme of the forum was: The Use of Quality Award Criteria and Models for Self-Assessment Purposes. A premise of the forum was "that for one company that applies for an award, perhaps one thousand companies use that award criteria and self-assessment model"

[1] The collective develops a set of common standards and common procedures for self-study and third-party review.

(Conti, 1994, p. 5). A recent survey by the American Society for Quality Control produced a similar conclusion regarding the Malcolm Baldrige National Quality Award criteria (Bemowski & Stratton, 1995). As was intended by the designers, quality award criteria and models are increasingly used for organizational self-assessments[2]. Little research has been found on the use of quality awards and models for self-assessment. "The time has come for a critical review of self-assessment criteria and methodologies" (Conti, 1994, p. 5). The proceedings of the Forum mentioned above were the best source of documentation found, yet the papers tended to be case studies or critical reviews. Deductive testing and empirical data appear scarce on this topic.

Organizational self-assessments using quality models[3] range from the simple to the elaborate. Gallagher (1994, pp. 93-95) described four approaches, presented here from the least to most formal: questionnaires or checklists; facilitator-led team; smaller written document or proformas; and award style self-assessments. Smith's (1994) research of the literature and commercial practices produced a taxonomy of five quality assessment approaches: questionnaire survey; focus group; audit; documented analysis; and continuous documentation. The focus of this research is on approaches like Smith's documented analysis or Gallagher's proformas and award style self-assessments, but the findings may have implications for other approaches as well.

Smith describes documented analysis as "a written report about the organization on each criterion and an evaluation of that report by independent 'examiners'" (1994, p. 185). Gallagher describes award style self-assessment as "a team is formed who set about preparing an award style submission document (up to 75 pages in length). A team of trained assessors then score the submission document determining strengths and areas for

[2] In fact, "the main purpose of the (Swedish Quality) Award program is that the guidelines should be used as a self-assessment tool" (Jernberg, Lindström, and Chocron, 1994, p.36).

[3] Quality models will be used here to refer to models that may or may not be part of an award process.

improvement for each criteria in the model" (1994, p. 95). Gallagher's proformas approach is similar to the award style, except the submission document is much shorter (e.g., 20 to 30 pages).

Gallagher's and Smith's questionnaire based approaches are essentially the same and self-explanatory. Smith's focus group approach is very similar to Ritter's (1993) description of Goal/QPC's use of facilitated small group discussion and self-scoring as a self-assessment tool. Gallagher's facilitator-led team and Smith's audit are essentially observation based self-assessments, with no formal document (self-study) preparation. Smith's continuous documentation approach is similar to documented analysis "except that the documentation is maintained as an on-going record" (1994, p. 187). This is a new approach which Smith intends to experiment with in the future.

The Malcolm Baldrige National Quality Award (MBNQA) criteria appear to be one of the most widely used standards for organizational self-assessments. The National Institute of Standards and Technology's (NIST) Baldrige program office maintains an incomplete listing of companies basing their internal quality awards[4] on the MBNQA criteria (Hertz, 1995). The list sounds like a Who's Who of the Fortune 500: Aeroquip, AT&T, Carrier, DuPont, Goodyear, IBM, Sprint, and Texaco. Evans (1994), Godfrey and Myers (1994), and Stundza (1992) identify other organizations using the MBNQA for self-assessment, some with and some without internal award programs. These organizations include: Aid Association for Lutherans, British Airways, Dow Chemical, Eastman Chemical, Federal Express, Intel, and Westinghouse. Even organizations outside the United States, like Telecom of Finland, are using the MBNQA criteria rather than the Finnish National Quality Award or European Quality Award because the MBNQA is seen

[4] Unlike Conti (1994), I consider internal quality awards to be improvement oriented rather than conformity oriented; therefore, I view self-assessment for internal quality awards in the context of self-regulation.

as a more mature system (Anttila, 1994). The President's Quality Award, the U.S. government's internal quality award, recently replaced its criteria[5] with those from the MBNQA (Lewis, 1994). This means that government agencies conducting self-assessments are likely to be using the MBNQA criteria in the future.

Selection and training of evaluators are frequently raised issues in the literature. Godfrey and Myers (1994) discuss the attributes that are considered in the selection of examiners for the AT&T Chairman's Quality Award. AT&T's corporate quality office (CQO) selects examiners to "represent a mix of operational and quality professionals. The CQO considers several attributes to achieve a balance between experience [sic] and new examiners, major functions (manufacturing, development, human resources, etc.), job levels and major types of organizations (business unit, operating division, corporate support function)" (Godfrey & Myers, 1994, p. 71).

In a subsequent paper, Myers and Heller (1995) cite diversity in business perspective and quality system perspective as desirable for AT&T's examiner teams, but fail to define these terms. British Airways uses teams of "managers, employees, and service partners" (Evans, 1994, p. 10) as assessors (evaluators) for departmental quality assessments[6]. Evans points out that smaller teams could be used, but "British Airways regards the assessment as an opportunity to involve and motivate employees in continuous improvement activities" (p. 10). KLM airlines follows a similar approach in using line managers for self-assessments or "line assessing line" (Gibson & Sluis, 1994, p. 119); however, when scoring the self-assessment reports, the scoring teams are supplemented by European Foundation for Quality Management (EFQM) trained assessors and external

[5] While the President's Quality Award criteria were said to be based upon the MBNQA criteria (Federal Quality Institute, 1994), they were structured quite differently. The new criteria will follow those of the MBNQA, with little or no modification.

[6] British Airways' Departmental Quality Assessments do not include independent third-party scoring. Scoring is done by a team including members from within and outside the unit of analysis.

consultants. The EFQM selects and trains the assessors for the European Quality Award[7]. "Assessors are either senior line managers or quality professionals and are selected so as to ensure appropriate representation from different countries and business sectors" (Gallagher, 1994, pp. 95-96).

Examiners for IBM's Market Driven Quality (MDQ) assessment process are selected based on experience and training, including: in-depth understanding of IBM MDQ strategy, knowledge of the application of the quality improvement methodologies and tools required to support the strategy (Benedetti & Bertorelli, 1994, p. 105). In addition to selecting evaluators based on characteristics similar to those described above, NYNEX recruits some evaluators who are examiners for local state (Massachusetts, Maine, and New York) quality awards (Smith, 1994).

Once selected, training of evaluators becomes an issue. Most formal self-assessment processes appear to include training of their evaluators: AT&T (Godfrey & Myers, 1994; Myers & Heller, 1995); Brasmotor Group (Dagnino & de Souza, 1994); British Airways (Evans, 1994); Baxter Healthcare (Sanford, 1993); EFQM model (Gallagher, 1994); IBM (Benedetti & Bertorelli, 1994); KLM (Gibson & Sluis, 1994); NYNEX, (Smith, 1994). Training for evaluators can range from rigorous multi-day workshops to almost no formal training[8].

The more involved training programs are typically modeled after those used for the major quality awards and include study of the assessment standards, the assessment

[7] While the European Quality Award is an *external* award process, its emphasis on self-assessment makes it appropriate for this discussion. According to Michael Gallagher (1994, p. 93), manager of the European Quality Award, "EFQM's key mission of promoting TQM in European Business is best served by the adoption of self-assessment as a routine part of normal business management. So much the better for the Awards process, if in their own good time, organisations practising self-assessment make an application for The Award."

[8] The earlier discussion of evaluator characteristics assumes that new evaluators have some knowledge of quality and productivity improvement concepts.

process, preparing feedback/reporting, and the scoring of case studies (Myers & Heller, 1995; O'Brien, 1994; Evans, 1994; Sanford, 1993; Gibson & Sluis, 1994; NIST, 1990). Where teams of evaluators are used, consensus scoring is often included in the training. Practice scoring of case studies is used to familiarize the evaluators with the performance associated with levels of scoring by comparing their scores to those of experts. At the other extreme, Bell and Wilson (1994) describe a small organization wishing to conduct a self-assessment that "could neither afford the time nor expense of having a member of staff trained as an assessor" (pp. 133-134). While the organization in their example obtained assistance from a local business school, Bell and Wilson cite the need for less expensive training options for small organizations. Jernberg, Lindström, and Chocron (1994) describe the use of university students as evaluators in Sweden. The students conduct both data collection and evaluation as part of their final papers (perhaps senior projects or masters projects). Jernberg et al. recognized the limited accuracy of this approach, but see it as one solution to the problem of limited resources.

Concern regarding reliability or efficacy of evaluators' scores was frequently mentioned in the papers of the First European Forum on Quality Self-Assessment[9] (Conti, 1994; Martellani, 1994; Jernberg, Lindström, & Chocron, 1994; Fuchs & Stuntebeck, 1994), but no data or analysis was provided. An example published in National Productivity Review might best illustrate this potential problem.

> During 1989 BQA [Baxter Quality Award] applicants were asked to submit self-assessment scores using the Baldrige criteria along with an application report ranging in length from fifty to seventy-five pages. Thirty-one applications were received and examined. The results were provocative. All the applicants scored themselves on a par with or superior to Baldrige winners. The formal examination results prepared by independent examiners, however, told a much different story, with scores in many cases hundreds of points lower. (Sanford, 1993, p. 39).

[9] Papers from four continents were submitted for publication prior to the forum, implying that the authors raised these concerns without the benefit of the interchange at the forum.

The large differences between the self-scores and the independent examiners' scores could be due to numerous factors. Bias on the part of the self-scorers[10] is an obvious choice. A leniency effect due to lack of knowledge and training on the criteria or scoring procedures is also a possible explanation. If the independent examiners' scores are accepted as estimates of true scores, then the self-scores would be said to be very inaccurate. While the Baxter Healthcare example is not surprising for a first year process, submitting inaccurate self-scores could be very embarrassing for the managers involved. IBM's MDQ Assessment process requests self-scores before an independent assessment team is asked to score the applicant (without benefit of the seeing the self-scores).

Accuracy is not the only issue raised in the literature. Consistency is also important. Fuchs and Stuntebeck (1994, p. 24) say of AT&T's self-assessment process, "it must assure to the greatest extent possible consistency among units being assessed and between subsequent assessments of the same unit." Similar concerns regarding consistency of self-assessment processes are raised by Martellani (1994), Bell and Wilson (1994), and Smith (1994). Consistency can even be a problem for trained evaluators. Raymond Wachniak, a senior examiner for the first year of the MBNQA and later a MBNQA judge, described between-examiner score ranges in excess of one hundred points on a thousand point scale for the scoring of written MBNQA applications (Wachniak, 1990). All of these examiners had been through the MBNQA examiner training.

Rater Error Research

While little research has been conducted regarding the scoring effectiveness[11] of self-assessment evaluators, the industrial and organizational (I/O) psychology literature

[10] Gibson and Sluis (1994, p. 128) cite one of KLM's lessons learned as "assessment done by (internal) outsiders avoids too positive a picture."

[11] Scoring effectiveness (or rating effectiveness) is used to describe any study where an index of the effectiveness of the scoring (or rating) is a dependent measure. Effectiveness implies that the scoring

describes considerable research regarding the psychometric qualities of rater data. The Journal of Applied Psychology and Organizational Behavior and Human Decision Processes frequently publish studies of the third-party ratings of individual performance. This research may have limited generalizability to the scoring of organizational self-assessments. The raters of individual performance could be viewed as analogous to evaluators. The ratings would be analogous to the scores of the self-assessment. The ratees would be analogous to the organizations. Whereas evaluators score written self-assessments, the raters score written vignettes of individual behavior, videotapes of individuals, or actual live performance. Reviewing the rater research provides ideas for studying the scoring effectiveness of evaluators.

Many early studies focused on issues related to rater error, "inadequacies of one sort or another in the ratings" (Saal, Downey, & Lahey, 1980, p. 413; also see Bernardin & Walter, 1977; Borman, 1975). Rater errors include: halo, "a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behavior" (Saal, Downey, & Lahey, 1980, p. 415), leniency or severity, a rater's tendency to assign average performance levels above or below the scale midpoints, and restriction of range, a rater's failure to discriminate among ratees (Saal, Downey, & Lahey, 1980; Bernardin & Walter, 1977). A fourth criterion used in rater error research, not defined as undesirable or erroneous, is interrater reliability or agreement, "the extent to which two or more raters independently provide similar ratings on given aspects of the same individuals' behaviors" (Saal, Downey, & Lahey, 1980, p. 419).

More recent research appears to question the use of rating error indices such as halo and leniency, favoring more direct measures of accuracy (Sulsky & Balzer, 1988;

procedure is doing what was expected. Examples might include studies to reduce rater error or to increase the accuracy of scores.

Stamoulis & Hauenstein, 1993). Accuracy of measurement describes both the correlation and distance between one set of measures and a corresponding set of measures (e.g., true scores) accepted as a standard for comparison. (Sulsky & Balzer, 1988) Furthermore, accuracy scores make no assumptions regarding the actual distribution of ratee performance (i.e., traditional rater error measures are computed under the assumption that performance ratings are normally distributed with zero correlations among rating dimensions (Schwab, Heneman, & DeCotiis, 1975).

Problems with the dependent measures chosen, sample selection, and training type limit the generalizability of the findings from the rater error research (Saal, Downey, & Lahey, 1980; Sulsky & Balzer, 1988). Sulsky and Balzer (1988) reviewed studies of performance rating accuracy and found weak relations among the different accuracy operational definitions. Accuracy measures used in the performance-rating research are often based on the squared difference between a rater's scores and the true scores averaged across dimensions and ratees - the D^2 index. These measures produce an index of accuracy for a single rater.

Cronbach (1955) argued that a single index of accuracy aggregated too much information and proposed decomposing the D^2 index into "four separable and conceptually independent component accuracy scores" (Sulsky & Balzer, 1988, pp. 498-499). Variations and subsets of Cronbach's four component scores were used frequently in subsequent research (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; McIntyre, Smith, & Hassett, 1984; Pulakos, 1986; Stamoulis & Hauenstein, 1993).

Hauenstein and Alexander (1991) proposed two measures of accuracy for a single ratee study based on the squared difference between subject ratings and true scores averaged across dimensions. The first measure, elevation (E), is similar to Cronbach's (1955) elevation component of accuracy. Elevation represents the difference between mean subject ratings and mean true scores. Sulsky and Balzer (1988) describe elevation as being analogous to the differential grand mean in an analysis of variance (ANOVA). Hauenstein and Alexander's second measure, dimensional accuracy (DA), is comparable to Cronbach's (1955) differential accuracy[12]. Dimensional accuracy measures the accuracy with which each rater evaluated a single ratee on each dimension. These terms can be better understood by describing them under conditions of perfect accuracy:

> For elevation, perfect accuracy requires a rater's average observed rating to equal the average of the target scores. Perfect accuracy in terms of dimensional accuracy requires[13] both a correlation of positive one between a rater's observed ratings and the target ratings, and that a rater's variance for his/her ratings equals the variance of the target scores (Hauenstein & Alexander, 1991, p. 308).

Murphy, Garcia, Kerkar, Martin, and Balzer (1982) note that the importance of a particular aspect of rating accuracy varies with the type of organizational decisions the rating is intended to support. Thus, training should be oriented toward the aspect of accuracy needed to support decision making. Sulsky and Balzer (1988) state that accuracy is important when cutoff scores are being used (e.g., for promotions or awards) and accuracy scores are also useful for evaluating the impact of rater training interventions.

A potential weakness of the rater research is the frequent use of undergraduate students as subjects (Smith, 1986). Since much of this research is conducted in psychology departments, the students are often recruited from the larger lower-level psychology courses, thus further limiting generalizability. Although the use of managers from real ongoing organizations as subjects is scarce[14], Smith's review did find "for the most part, training effects for non-student raters parallel those of student raters" (p. 38).

[12] Murphy, Garcia, Kerkar, Martin, and Balzer (1982, p. 321) state "differential accuracy is the component that corresponds most closely to the lay notion of accuracy or interpersonal sensitivity."

[13] Close examination of the formula for dimensional accuracy finds that a correlation of positive one between a rater's observed ratings and the target ratings and that a rater's variance for his/her ratings equals the variance of the target scores *is sufficient, but not required* for a perfect DA score.

[14] Borman (1975) is a frequently cited exception; however, Borman's study did not include a control group. Borman used Navy officers as subjects.

Differences in the content of the training provided can make comparisons between studies difficult. Training content typically follows one of two models: rater error (RE) training or frame-of-reference (FOR) (Bernardin & Walter, 1977; McIntyre, Smith, & Hassett, 1984; Stamoulis & Hauenstein, 1993). In rater error training, "raters are given training on common psychometric errors (e.g., leniency/severity, halo, and central tendency) and then are admonished to avoid them" (McIntyre, Smith, & Hassett, 1984, p. 147). Frame-of-reference training "is designed to 'tune raters' to a common frame of reference so that worker behaviors can be similarly assessed by different raters" (McIntyre, Smith, and Hassett, 1984, p. 147). "A frame of reference is achieved by presenting samples of job performance to trainees along with the appropriate or 'true' ratings assigned to the performance by trained experts" (Smith, 1986, p. 31). Smith defined a third content area, Performance Dimensions Training (PDimT), which "familiarizes the raters with the dimensions by which the performance is rated" (p. 30). Elements of Smith's PDimT appear in others' descriptions of RE and FOR training, and few studies were found that presented PDimT content in isolation.

Rater training generally improves one or more aspects of rater effectiveness, but may result in degradation of other aspects. McIntyre, Smith, and Hassett (1984) found FOR training to improve accuracy more than RE or no training, but no difference was found among training conditions for reducing leniency/severity errors. Borman (1975) found RE training to reduce halo error, but interrater reliability was lower after training. Bernardin and Pence (1980) found RE training to reduce leniency and halo error more than rater accuracy training[15] (RAT) or no training, but the RE group's ratings were significantly less accurate than the ratings of the other two groups. Stamoulis and

[15] RAT was unique to Bernardin and Pence's study, although Smith categorized it as PDimT. Bernardin and Pence found no significant differences between the RAT group and the control group ratings.

Hauenstein (1993) found FOR training superior for improving dimensional accuracy, but RET trainees showed the most improvement for elevation accuracy. Their "results also suggested that increased variability in observed ratings led to RET's positive effects of elevation accuracy" (p. 1002). This increased variability could be problematic in a situation where only a few raters are being used. Pulakos (1986, p. 90) concluded that "there is no one 'best' way to train raters to provide accurate performance ratings." Smith (1986) concluded that "the more actively involved the raters become in the training process, the greater the outcome. . . . The evidence suggests that the best way to increase accuracy is to combine the two [FOR and PDimT] training approaches."

Strategic Quality Management

Awareness of the importance of quality in the products, services, and processes of organizations grew quickly during the 1980s. Garvin's (1988) description of the evolution from inspection, statistical quality control, and quality assurance to strategic quality management summarizes much of what has been written about the quality movement. Deming (1986) long argued that improving quality led to improved productivity and, thus enhanced competitive position. Review of the professional and business journals of the 1980's clearly demonstrates the increased awareness of the importance of quality. But the question of why quickly turned to what to improve and how.

The demand for ever-improving levels of quality led to searches for methods, approaches, models, and recipes for quality improvement (Business Week, 1991; Deming, 1986, 1993). Experts such as Deming, Juran (1986), and Crosby (1979) were widely followed in efforts to learn the path to quality. Articles describing the latest uses of individual quality tools were published in the professional and business journals (e.g., see

Quality Progress, Industrial Engineering[16], Harvard Business Review). Consulting firms offered packages of tools tied together by their model of strategic quality management (e.g., see the catalogs of the firms listed on the Federal Supply Schedule for Total Quality Management). Each expert, tool, and model had its advocates. Differences in philosophy and language made finding common ground difficult.

The Malcolm Baldrige National Quality Award

The Malcolm Baldrige National Quality Award (MBNQA) was established in 1988 by Public Law 100-107 to improve quality and productivity by:

> A. helping to stimulate American companies to improve quality and productivity for the pride of recognition while obtaining a competitive edge through increased profits;
>
> B. recognizing the achievements of those companies which improve the quality of their goods and services and providing an example to others;
>
> C. establishing guidelines and criteria that can be used by business, industrial, governmental, and other organizations in evaluating their own quality improvement efforts, and
>
> D. providing specific guidance for other American organizations that wish to learn how to manage for high quality by making available detailed information on how winning organizations were able to change their cultures and achieve eminence. (NIST, 1989)

The MBNQA continues to evolve since it was established (Stratton, 1990; Reimann, 1992; NIST, 1994a). The award criteria are reviewed and revised each year, but their basic purposes remain consistent with Public Law 100-107. The 1995 award criteria comprise seven categories built upon the following core values and concepts: customer-driven quality, leadership, continuous improvement and learning, employee participation and development, fast response, design quality and prevention, long-range view of the future, management by fact, partnership development, corporate responsibility and citizenship, and

[16] Now known as IIE Solutions.

results orientation (NIST, 1994a). The Baldrige Award criteria framework or model illustrates the relations between the seven categories (see Appendix A). The seven categories are:

> 1.0 Leadership;
> 2.0 Information and Analysis;
> 3.0 Strategic Planning;
> 4.0 Human Resource Development and Management;
> 5.0 Process Management;
> 6.0 Business Results;
> 7.0 Customer Focus and Satisfaction. (NIST, 1994a)

The seven categories are comprised of a total of twenty-four items and it is at the item level that scoring occurs (see Appendix B for a listing of items by category).

The award process is similar to that of other quality awards and easily fits the self-assessment and validation components of the self-regulation model presented in Chapter I. The process follows these basic steps:

> 1. The applicant organization prepares an application package based upon the award criteria (i.e., a structured self-study),
>
> 2. The application package is submitted to the Board of Examiners for evaluation (i.e., third-party review and scoring),
>
> 3. The Board of Examiners decides which applicants have scored high enough to be considered for the award and conducts a site visit to clarify and verify any outstanding issues (see Appendix C for a flow chart of the evaluation process),
>
> 4. The Award Judges recommend recipients of the Baldrige Award to the Secretary of Commerce (i.e., final validation of scoring). (NIST, 1995a)

The actual scoring of each item is done on a 0 to 100% scale, using 10% increments based on three evaluation dimensions: (1) Approach; (2) Deployment; and (3) Results.

> "Approach" refers to how the applicant addresses the Item requirements - the <u>method</u>(s) used. . . "Deployment" refers to the <u>extent</u> to which the applicant's approach is applied to all requirements of the Item. . . "Results" refers to <u>outcomes</u> in achieving the purposes given in the Item.
>
> Award Examination Items are classified according to the kinds of information and/or data applicants are expected to furnish. The two types of Items and their designations are:

1. Approach/Deployment
2. Results (NIST, 1994a).

This means each item is scored on a scale based on Approach/Deployment or on a scale based on Results. Each item is scored as one or the other, no item is scored using both even though the item content may reflect all three dimensions. The scoring guidelines used for each classification are shown in Appendix D.

The MBNQA has become widely accepted. Approximately 1,000,000 copies of the award criteria have been requested, yet only 546 organizations have applied for the award (Bemowski & Stratton, 1995). Bemowski and Stratton sampled recent recipients of the criteria and found that between 35 and 44% of their respondents used the criteria for either a department-wide or company-wide self-assessment. The MBNQA is not without criticism (see Pyzdek, 1995; Bowles, 1992; Crosby, 1992; Deming, 1992). Review of the proponents and critics (Harvard Business Review, 1992) supports Garvin's (1992, p. 84) proposition that the award sits firmly between the two extremes of a narrowly defined award for product and service excellence and an all-encompassing award for overall management excellence.

Other Applications of Self-assessment

Academic accreditation is one of the most commonly practiced uses of self-assessment. The educational literature abounds with writings describing the self-assessment and validation components of accreditation (Craft, 1992; El-Khawas, 1983; Kells, 1992a; Petersen, 1979; Selden and Porter, 1977; Young, Chambers, Kells, and Associates, 1983). Higher education has been practicing self-study and third-party review for decades; unfortunately, the linkage to a continual process of improvement is often missing[17].

Academic accreditation is often perceived as conformity self-assessment; however, most accreditation processes cite improvement as a primary purpose (ABET, 1994; Kells, 1992a, Young et al., 1983). But there is little mention of improvement as a process or how improvement is linked to the planning, implementing, and measuring of continual improvement. The Accreditation Board for Engineering and Technology (ABET) has recognized this shortcoming in their accreditation systems. ABET has proposed "revolutionary criteria" to support a "quality-oriented, innovation supportive accreditation system" (Parrish, 1994, p. 16). Examination of ABET's proposed Revised Engineering Criteria (ABET, 1995) shows the need for qualified and trained evaluators will be even greater than before (Pape, 1995). The new criteria are more concise and less prescriptive than those they replace. The ability to make consistent decisions across evaluators may be key to the success of the new criteria.

Traditional conformance audits can be transformed into a form of self-regulation, possibly serving both conformance and improvement purposes. Aquino (1990) described the used of self-assessments at Westinghouse Electronic Systems Group as being near the middle of the review spectrum. "At one end of the spectrum is the traditional audit review aimed solely at compliance. At the opposite end is a review aimed solely as improvement. In between are types of reviews that have a mix of these two objectives - for instance, an assessment of a key supplier by a customer, where both have been working together to improve quality" (Aquino, 1990, p. 48).

[17] Petersen (1979) found little mention of improvement *processes* in her meta-analysis of accreditation standards.

The U.S. Air Force's Air Mobility Command (AMC) changed the name (and mission) of its Inspector General organization to Quality Support and Readiness (QS). Rather than conducting the typical policy and procedures compliance visit, the QS organization conducts quality visits. Quality visits serve the role of third-party review, validating the self-assessments of the visited unit. Quality visits also include a team who helps the visited organization with any process problems (Bemowski, 1992).

Talley (1989) described early efforts by the national contractor accreditation system (NCAS) to provide "a cost effective means for supplier quality system accreditation and product qualification." Using a third party accreditation system reduced surveillance costs and increased standardization amongst suppliers. Pilot organizations were suppliers to defense prime contractors, including laboratory services. NCAS used this third party accreditation to replace an attempt at solving the counterfeit fasteners problem by legislation (i.e., used collective self-regulation in lieu of regulatory compliance).

III. Research Methodology

The Research Problem Revisited and Preview of the Experimental Design

This research examined the impact of evaluator training on the consistency and accuracy of third-party scoring of organizational self-assessments and the relationship of evaluator characteristics to the accuracy of the third-party scoring. The study also demonstrated the magnitude of score variation and error possible for a sample of novice evaluators using an established scoring system.

The study was conducted by having evaluators score a written organizational self-assessment using the criteria and scoring system of the Malcolm Baldrige National Quality Award[1]. A classic experimental design was used. Each evaluator scored two (out of seven) randomly assigned categories[2] prior to the training intervention (i.e., treatment). The evaluators assigned to the control group then scored two additional randomly assigned categories. Finally, the evaluators assigned to the treatment group were given a modest training program and then scored their two additional randomly assigned categories. Scores and supporting comments were compiled and analyzed. Measures of consistency and accuracy were calculated, examined, and compared between groups and evaluations (i.e., first versus second). Analyses were performed on these data to address the research questions and hypotheses described below.

Consistency is the degree of agreement between evaluators (adapted from Bernardin & Walter, 1977). Consistency may also be viewed as the relative lack of variation between evaluators. Accuracy is measured as a function of the relative distance between an

[1] The Baldrige Award was chosen because of the wide acceptance of its criteria and scoring system. Also, the Baldrige Award annually produces a case study (in the form of written organizational self-assessments) for the training of its examiners. Evaluation materials, including the scores and supporting comments from a team of experts (i.e., Senior Examiners) who evaluated the case study are also provided.

[2] Evaluators were assigned to categories such that approximately two-sevenths of the evaluators were assigned to each category for the initial evaluation. Another two-sevenths was assigned to each category for the second evaluation.

evaluator's scores and the true scores of ratee performance (adapted from Sulsky & Balzer, 1988). Accuracy may also be viewed as the relative absence of error in the scores of a given evaluator. Consistency addresses relative agreement among the scores of two or more evaluators[3], while accuracy addresses variation between the scores of an individual evaluator and the true scores.

Research Questions

A number of research questions were developed to refine the focus of the research. The research problem is too broad to be addressed by a single question, but each of these research questions contributes to addressing a particular facet of the research problem. The experiment provided evidence to address each of these questions. The research hypotheses, data needed, and data analyses for each question[4] are described in the following section. The actual data collected and the specific hypotheses tested for the analyses are presented in Chapter IV. To assist in tracking the questions throughout this document, each question has been given a number preceded by the letter "Q."

The initial set of research questions relate to the scores given a particular ratee by a sample of relatively untrained[5] evaluators. The first questions examine the consistency of the evaluators' scores, followed by questions regarding the accuracy of the evaluators' scores. Consistency is the degree of agreement between evaluators. Questions one and two looked at the consistency of scores between evaluators who are scoring the same item or category.

> Q1: How much agreement is there among evaluators on the score of an item?

[3] Or the relative agreement between two or more groups of evaluators.

[4] The reader may wish to refer to Table 1 as this section is read. Table 1 summarizes the research questions, research hypotheses, data needs, and data analyses.

[5] Meaning they have been given no explicit training for using the MBNQA criteria to score organizational self-assessments. They have been given the Award Criteria booklet well in advance and been asked to read and familiarize themselves with the criteria and scoring system.

Q2: How much agreement is there among evaluators on the score of a category?

Questions one and two were addressed for each item and each category of the MBNQA criteria. The results were expected to demonstrate how much agreement might be seen, for each item and category, across a sample of untrained evaluators.

Questions three and four examine the consistency of score variation between items and categories. Question three examines the consistency of score variation between the items of a category and was addressed for each category. Question four examines the consistency of score variation across all seven categories.

Q3: How consistent is the within-item variation of evaluator scores across all the items of a category?

Q4: How consistent is the within-category variation of evaluator scores across all seven categories?

Addressing these questions was intended to identify particular items or categories that appear more susceptible to variation in scores.

Question five (Q5) addresses the issue of score accuracy for a sample of untrained evaluators. Accuracy indicates how close an evaluator's scores are to the right answer; however, distance alone only gives a partial view of accuracy. Correlation and relative variance across dimensions are also important to accuracy. Accuracy measures used in the performance-rating research are often based on the squared difference between a rater's scores and the true scores averaged across dimensions and ratees, the D^2 index (Sulsky and Balzer, 1988).

The D^2 index can be decomposed into "four separable and conceptually independent component accuracy scores" (Sulsky and Balzer, 1988, pp. 498-499). Hauenstein and Alexander (1991, p. 308) point out that only two of these four "components accuracy can be computed in a single ratee study." For Q5, Hauenstein and

Alexander's two measures of accuracy were used: Elevation[6] (E) and Dimensional Accuracy[7] (DA). These terms are operationally defined below, adapted from Hauenstein and Alexander (1991):

$$E = \sqrt{(\bar{x}_{.} - \bar{t}_{.})^2}$$

where $\bar{x}_{.}$ and $\bar{t}_{.}$ = observed mean rating and mean target score over all dimensions

$$DA = \sqrt{1/n\sum[(x_j - \bar{x}_{.}) - (t_j - \bar{t}_{.})]^2}$$

where x_j and t_j = observed rating and target score on dimension j, and $\bar{x}_{.}$ and $\bar{t}_{.}$ are defined as above.

Elevation and Dimensional Accuracy are illustrated graphically for a hypothetical category with three items in Appendix AT. Additional illustrations are included in Chapter V with the discussion of results related to accuracy. For Q5, individual and mean accuracy indices (i.e., E and DA) were calculated for each category.

Q5: How accurate are the evaluators' scores for each category?

Answering questions one through five was intended to describe the consistency and accuracy of the scores of untrained evaluators, and to provide a baseline for comparison with the results of the second evaluations. Approximately one-half of the evaluators (the control group) conducted a second evaluation (scoring) immediately before[8] the training intervention and the other half (the treatment group) conducted a second evaluation

[6] Elevation is the average of J's predictions over all items and ratees minus the central tendency of the self-descriptions ("true" scores) for all items and ratees combined (Cronbach, 1955, p. 178). Sulsky & Balzer's (1988) two-way ANOVA analogy described Elevation as the difference between the rater's grand mean and the true score grand mean. Elevation is the mean difference between the evaluator's scores and the true scores.

[7] Dimensional accuracy (DA) is "equivalent to Cronbach's (1955) differential accuracy in that it measures the accuracy with which each rater evaluated a single ratee on each dimension. However, because there was only one ratee, we chose not use the label 'differential' accuracy" (Hauenstein and Alexander, 1991, p. 308). Dimensional accuracy is similar to McIntyre, Smith, and Hassett's correlational accuracy, in that it "measures the parallelism between subjects' scores and true scores" (1984, p. 151).

[8] This allowed the subjects in the control group to attend the training intervention without affecting their second evaluations.

immediately after the training intervention. The same ratee was evaluated in the second evaluation, but each evaluator evaluated a different pair[9] of categories.

Questions six through ten address the comparison of the scores from the initial evaluation (pre-training) to the scores from the second evaluation (post-training), using the control group to account for any learning effect. Questions six through ten correspond to the same issues as questions one through five, but the focus is on change rather than absolute levels of consistency and accuracy. Again, questions six through ten apply to a given sample of originally untrained evaluators; however, the treatment group received training between the first and second evaluations.

Q6: Did agreement among evaluators on the score of an item change (improve) due to evaluator training?

Q7: Did agreement among evaluators on the score of a category change (improve) due to evaluator training?

Questions six and seven were answered for each item and each category of the MBNQA criteria. These questions were addressed from both a statistical and practical perspective. For example, if the agreement among evaluators were shown to change by a statistically significant amount, then the questions become: How much was this change in terms of raw scores? Was there a difference between the magnitude of change exhibited by the control group versus that exhibited by the treatment group?

Questions eight and nine examine changes in the consistency of score variation.

Q8: Did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?

Q9: Did within-category variation of evaluator scores across all seven categories change (decrease) due to evaluator training?

[9] To present the subjects with a manageable task load, each subject was only asked to evaluate two randomly assigned categories per evaluation. That is, each category was evaluated by two-sevenths of all the subjects during both the first and second evaluations. After both evaluations, each subject had evaluated four randomly assigned categories.

Questions eight and nine were partially addressed by questions six and seven. Whereas questions six and seven looked for reduction in variation for a given item or category, questions eight and nine looked for patterns of reduction across items or categories. Question nine also addressed any change in the relative variation between categories.

Question ten addresses the issue of improvement in score accuracy due to the training intervention. Again, both measures of elevation and dimensional accuracy were used.

Q10: Did the accuracy of the evaluators' scores change (improve) due to the evaluator training?

Question ten was addressed for each category. Change due to time (first versus second evaluation) and change due to group (control versus treatment) were examined.

Evaluator training was an important component of this research project. The data generated were intended to provide insights into how training might be improved. NIST identifies potential improvements to the case studies it uses for training MBNQA examiners by studying the range and standard deviation of the scores examiners give the case study. Categories or items with wider variation are seen as areas where the training made need strengthening. NIST believes the variation is smaller for experienced versus new examiners, but has performed only rudimentary analyses of the data. None of these analyses have been published. In fact, NIST destroys these data once they have been used to identify improvements to examiner training (Hertz, 1995). Question 11 does not directly address the research problem, but addresses an obvious question for this research project.

Q11: How might the training of evaluators be improved?

Depending on the results of the experiment, the answer to Q11 could be very important. If training does not appear to have an effect on the accuracy of evaluator scores, then the adequacy of the training comes into question. If training does appear to have an effect on the accuracy of evaluator scores, then the question of how to amplify the effect arises.

The final research question addresses the relationship between evaluator characteristics and the accuracy of their scores.

Q12: Which rater characteristics best predict the accuracy of the evaluators' scores?

Characteristics identified as important for the selection of evaluators[10] were measured for each subject and compared to the accuracy of their scores. These data were explored and the results used to propose a model of the relationships between selected characteristics and the predicted accuracy of evaluator scores.

Research Hypotheses

The research questions in the previous section were used to develop research hypotheses[11]. Not every research question lent itself to the development of hypotheses. Such questions were addressed by direct analysis and interpretation of the data. Other questions were addressed by developing and testing one or more hypotheses. Still others were addressed by a combination of these approaches. The following paragraphs list each research question, identifying those that were addressed by hypothesis testing and describing each research hypothesis relevant to the question. The specific testable hypotheses for each research hypothesis are presented in Chapter IV. The data needs and the data analyses to be performed are identified for each question and its supporting research hypotheses (if applicable). Table 1 summarizes how the research problem is broken into research questions with corresponding research hypotheses, data needs, and the data analyses to be performed.

[10] These characteristics have been identified by comparing and contrasting lists of desirable characteristics espoused by a sample of organizations conducting third-party scoring of organizational self-assessments.

[11] The research hypotheses were not written in a testable form, but in a form that illustrated the researcher's expectations and facilitated the development of testable hypotheses. In some cases, a single research hypothesis required developing and testing several testable hypotheses. The testable hypotheses are presented in Chapter IV.

Hypothesis one (H1) addresses the first set of research questions[12] related to consistency, rather than addressing a specific research question. This hypothesis was intended to establish the baseline condition, that there were no differences between the scores of the treatment and control groups. Testable hypotheses for H1 were developed and tested at both the item and category levels.

H1: a) There will be no difference in item scores between the treatment and control groups during the initial evaluation.

b) There will be no difference in category scores between the treatment and control groups during the initial evaluation.

H1 required collecting the evaluators' (initial evaluation) scores for each item and computing a category score based on the average of the category's item scores[13]. Testable hypotheses were developed and t-tests were performed on the group mean scores for each item and each category. Descriptive statistics, including mean, standard deviation, range, and interquartile range were calculated for each group on each item and category. Boxplots of group item and category scores were contrasted.

Question 1 asked "how much agreement is there among evaluators on the score of an item?" Hypothesis 2 (H2) addressed this question and led to the development of testable hypotheses. These hypotheses tested for differences in item score variation between the treatment and control groups. Comparisons of descriptive statistics and boxplots were also used to address Questions 1 and 2.

H2: For each item, there will be no difference in score variances between the treatment and control groups.

[12] A broad question was later identified as an "umbrella" for this first set of questions and H1 may be viewed as partially addressing this broad question.

[13] All items are scored on a 0 to 100% scale, in 10% increments. The item scores are differentially weighted when calculating the point value of a category for the Baldrige Award; however, this weighting will not be used for the statistical analysis. The differing weights of items and categories will be considered when identifying the practical implications of the statistical analyses.

Table 1. Structure of the Research Problem

Research Questions	Hypotheses	Data[14] Needed	Data Analyses
For a given sample of untrained evaluators, what is the consistency of their scores?	H1: There will be no difference in scores between the treatment and control groups during the initial evaluation: - a) by item, and - b) by category.	Each evaluator's initial evaluation scores for two categories of the case study (categories randomly assigned to evaluators).	- t-tests of group mean scores for each item and each category (H1); - descriptive statistics[15] and a boxplot of scores for each item and category.
Q1: How much agreement is there among evaluators on the score of an item?	H2: For each item, there will be no difference in score variances between the treatment and control groups.		- F-tests of item score variances between treatment and control groups (H2).
Q2: How much agreement is there among evaluators on the score of a category?	H3: For each category, there will be no difference in score variances between the treatment and control groups.		- F-tests of category score variances between treatment and control groups (H3).
Q3: How consistent is the within-item variation of evaluator scores across all the items of a category?	- na		- compare boxplots across items of a category.
Q4: How consistent is the within-category variation of evaluator scores across all seven categories?	H4: There will be a difference in score variances between categories.		- compare boxplots of category scores across all seven categories; - Hartley's Fmax test for homogeneity of population variances (H4), followed by pairwise comparisons of all 21 pairs (dropping data points for those subjects who evaluated both categories of the pair) using a Bonferroni adjustment to control alpha.

table continues

[14] A data need will not necessarily be repeated once it has been identified; otherwise, this column would be very repetitious.

[15] Descriptive statistics include mean, standard deviation, range, and interquartile range.

Research Questions	Hypotheses	Data Needed	Data Analyses
For a given sample of untrained evaluators, what is the accuracy of their scores? Q5: How accurate are the evaluators' scores for each category?	H5: There will be no difference in accuracy between the treatment and control groups during their first evaluations:	Each evaluator's initial evaluation scores for two categories of the case study.	- plot expert scores onto boxplots of scores for each item and category; - calculate Elevation[16] (E) and Dimensional accuracy[17] (DA) for each evaluator for each category, plot to test for normality; t-tests of Group mean Elevation for each category, t-tests of Group mean DA for each category (H5).
		Each evaluator's initial evaluation comments on strengths, areas for improvement, and site visit issues	- qualitative analyses, compare content of evaluators' comments to the experts' comments

table continues

[16] Elevation is the average of J's predictions over all items and ratees minus the central tendency of the self-descriptions ("true" scores) for all items and ratees combined (Cronbach, 1955, p. 178). Sulsky & Balzer's (1988) two-way ANOVA analogy described Elevation as the difference between the rater's grand mean and the true score grand mean.

[17] Dimensional accuracy (DA) is "equivalent to Cronbach's (1955) differential accuracy in that it measures the accuracy with which each rater evaluated a single ratee on each dimension" (Hauenstein and Alexander, 1991, p. 308). Dimensional accuracy is similar to McIntyre, Smith, and Hassett's correlational accuracy, in that it "measures the parallelism between subjects' scores and true scores" (1984, p. 151).

Research Questions	Hypotheses	Data Needed	Data Analyses
For a given sample of untrained evaluators, will evaluator training change the consistency of their scores?	H6: There will be a difference in scores between the treatment and control groups during the second evaluation: - a) by item, and - b) by category.	Each evaluator's second evaluation scores for two categories of the case study (subjects randomly assigned to categories, without repeating the same categories).	- descriptive statistics and a boxplot of scores (overlay on initial evaluation boxplots) for each item and category. - two-way ANOVAs of group X time for each item and category with evaluator scores as data points (H6).
Q6: Did agreement among evaluators on the score of an item change (improve) due to evaluator training?	H7: Item score variances will be smaller for second evaluation scores than for first evaluation scores.		- F-tests of item score variances between initial and second evaluations for each group, and between treatment and control groups' second evaluations. (H7) Similar F-tests for each category. (H8)
Q7: Did agreement among evaluators on the score of a category change (improve) due to evaluator training?	H8: Category score variances will be smaller for second evaluation scores than for first evaluation scores.		
Q8: Did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?	- na		- compare initial and second evaluation boxplots of item scores for all the items in a category.
Q9: Did within-category variation of evaluator scores across all seven categories change (decrease) due to evaluator training?	-H9: There will be a difference in score variances between categories for both the treatment and control groups.		- compare initial and second evaluation boxplots of category scores across all seven categories; - Hartley's Fmax test (H9), followed by comparisons of all 21 pairs (dropping data points for those subjects who evaluated both categories of the pair) using a Bonferroni adjustment to control alpha.

table continues

Research Questions	Hypotheses	Data Needed	Data Analyses
For a given sample of untrained evaluators, how will evaluator training change the accuracy of their scores? Q10: Did the accuracy of the evaluators' scores change (improve) due to the evaluator training?	H10: The accuracy of evaluators' scores will improve between the first and second evaluations.	Each evaluator's second evaluation scores for two categories of the case study. Each evaluator's second evaluation comments on strengths, areas for improvement, and site visit issues	- plot expert scores onto boxplots of scores for each item and category, compare with initial boxplots - calculate Elevation (E) and Dimensional Accuracy (DA) for each evaluator for each category, plot to test for normality; - two-way ANOVAs of group X time for each category with evaluator DA as data points (H10); - two-way ANOVAs of group X time for each category with evaluator E as data points (H10); - qualitative analyses, compare content of evaluators' comments to the experts' comments
Q11: How might the training of evaluators be improved?	H11: Subjects' perceived difficulty of evaluating a category will be negatively related to their accuracy in evaluating that category. H12: Subjects' perceived accuracy in scoring a category will be positively related to their accuracy in evaluating that category.	Subjects' perception of the difficulty of evaluating each category (each subject will rate 4 categories). Subjects' perception of the accuracy of scoring each category.	- For each category, comparison of subjects' perceived difficulty to level of DA (test for a negative relation using an ordered categories method); repeat using elevation rather than DA. (H11) - For each category, comparison of subjects' perceived accuracy to level of DA. (test for a positive relation using an ordered categories method); repeat using elevation rather than DA. (H12)

table continues

Research Questions	Hypotheses	Data Needed	Data Analyses
Q12: Which (of the following) evaluator characteristics best predict the accuracy of the evaluators' scores? - previous use of MB criteria - level of education - educational specialty - amount of Q&P training - work experience - job function - job level - employer industry - employer size - age - gender	- na	Subjects' responses to a survey of demographic characteristics (collected during second evaluations).	- Descriptive statistics for each characteristic by group and combined; - Step-wise multiple regression to explore the data.

Variances were calculated for the scores of each item and category. Hypotheses were tested for each item by conducting F-tests of item score variation between the treatment and control groups.

Question 2 asked "how much agreement is there among evaluators on the score of a category?" Hypothesis 3 (H3) addressed this question and led to the development of testable hypotheses. These hypotheses tested for differences in category score variances between the treatment and control groups.

H3: For each category, there will be no difference in score variances between the treatment and control groups.

Hypotheses were tested for each category by conducting F-tests of category score variances between the treatment and control groups.

Question 3 was not addressed by hypothesis testing. Question 3 asked "how consistent is the within-item variation of evaluator scores across all the items of a category?" Since H2 implied no differences in item score variances between the treatment and control groups, the group scores were pooled for this analysis. Boxplots of the scores for each item were contrasted on a category by category basis. That is, boxplots of all the items of a category were plotted at the same scale to facilitate comparison.

Question 4 asked "how consistent is the within-category variation of evaluator scores across all seven categories?" The hypotheses developed under Hypothesis 4 (H4) addressed this question by testing for differences in the category score variances across all seven categories. Comparing boxplots of category scores across all seven categories was used to supplement the results of testing the hypotheses under H4.

H4: There will be a difference in score variances between categories.

Since H3 implied no differences in category score variances between the treatment and control groups, the group scores were pooled for this analysis. Hartley's F_{max} test (Ott, 1984) was used to test for homogeneity of population variances. Assuming the F_{max} test finds a difference in score variances between categories, the question of which categories exhibit more variance will be addressed by pairwise comparisons of the twenty-one possible pairs of categories. Two adjustments are necessary for these pairwise comparisons. First, the scores of those evaluators who evaluated both the categories in a pair must be dropped due to dependence considerations. Second, a Bonferroni adjustment (Hays, 1988) should be made to reduce the likelihood of a Type-I error.

Question 5 asked "how accurate are the evaluators' scores for each category?" Q5 was addressed by graphical analysis and hypothesis testing. Expert scores for each item and category were plotted onto the boxplots of evaluator scores described earlier. This facilitated comparison of expert scores to the central tendency of the evaluators' scores. Hypotheses developed under Hypothesis 5 (H5) tested for differences in accuracy between the treatment and control groups.

> H5: There will be no difference in accuracy between the treatment and control groups during their first evaluations.

Elevation and dimensional accuracy were calculated for each evaluator for each category. Assuming the elevation and DA indices for a category appear normal, t-tests of group mean elevation and t-tests of group mean DA would be conducted for each category. If not, a suitable nonparametric or rank-based procedure would be used to test group medians. A content analysis approach was used to compare the evaluators' comments to the experts' comments[18] on strengths, areas for improvement, and site visit issues.

Questions six through ten addressed the comparison of the scores from the initial evaluation (pre-training) to the scores from the second evaluation (post-training) using the control group to account for any learning effect. Hypothesis 6 (H6) addressed the overall issue of change in scores and doesn't apply to one specific research question[19].

[18] These are the qualitative comments an evaluator provides for each item to justify his or her score and to provide feedback to the organization being evaluated.

[19] A broad question was later identified as an "umbrella" for this set of questions (see Table 1) and H6 may be viewed as partially addressing this broad question.

H6: a) There will be a difference in item scores between the treatment and control groups during the second evaluation.

b) There will be a difference in category scores between the treatment and control groups during the second evaluation.

Addressing H6 required collecting the evaluators' (second evaluation) scores for each item and computing category scores[20] for each evaluator. Group X time analyses of variance (ANOVA) were conducted for each item and each category, using evaluator scores as data points. Descriptive statistics, including mean, standard deviation, range, and interquartile range were calculated for each group on each item and category. Boxplots of group item and category scores for both the first and second evaluation were contrasted (i.e., for a given item or category, boxplots of first evaluation treatment group scores, first evaluation control groups scores, second evaluation treatment group scores, and second evaluation control group scores were compared).

Question 6 asked "did agreement among evaluators on the score of an item change (improve) due to evaluator training?" The hypotheses developed under Hypothesis 7 (H7) addressed this by testing for reduction of item score variances.

H7: Item score variances will be smaller for second evaluation scores than for first evaluation scores.

Variances were calculated for the (second evaluation) scores of each item and category. Hypotheses were tested for each item by conducting F-tests of item score variance between the first and second evaluations for each group. F-tests were also used to test for differences between the variances of the groups' second evaluation scores.

Question 7 asked "did agreement among evaluators on the score of a category change (improve) due to evaluator training?" The testable hypotheses developed under Hypothesis 8 (H8) addressed this question by testing for reduction of category score variances.

[20] Category scores are based on the simple average of the category's item scores.

H8: Category score variances will be smaller for second evaluation scores than for first evaluation scores.

The hypotheses under H8 were tested for each category by conducting F-tests of category score variances between the first and second evaluations for each group. The difference between second evaluation score variances of the groups was also tested.

Question 8 asked "did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?" The results of testing the hypotheses under H7, for all the items in each category, were used to address Question 8. Comparison of item score boxplots on a category by category basis was also used to address Q8.

Question 9 asked "did within-category variation of evaluator scores across all seven categories change (decrease) due to evaluator training?" The testable hypotheses developed under Hypothesis 9 (H9) addressed this question by testing for differences in the category score variances across all seven categories.

H9: a) There will be a difference in score variances between categories for the treatment group.

b) There will be a difference in score variances between categories for the control group.

Q9 was also addressed by comparing initial and second evaluation boxplots of group category scores across all seven categories. The hypotheses developed under H9 were tested using Hartley's F_{max} test for homogeneity of population variances. The results from H9 were compared to the results from H4 to see if the training intervention appeared to affect the relative score variance across categories. As with H4, if the F_{max} test found a difference in score variance between categories, pairwise comparisons will be used to identify which categories exhibit more variance.

Question 10 asked "did the accuracy of the evaluators' scores change (improve) due to the evaluator training?" This was addressed by graphical analysis and hypothesis testing. Initial and second evaluation boxplots for each item and category, with the experts' scores overplotted, were compared. The hypotheses developed under Hypothesis 10 (H10) tested for differences in accuracy due to the interaction of time (first or second evaluation) and group (treatment or control).

> H10: The accuracy of evaluators' scores will improve between the first and second evaluations.

Elevation and dimensional accuracy were calculated for each evaluator that scored each category. Assuming the elevation and DA indices for a category appear normal, two-way ANOVAs of group X time would be conducted for each category with evaluator elevation as data points. The same two-way ANOVAs would be repeated with evaluator DA as data points. If the plots of elevation or DA did not appear normal, a suitable nonparametric or rank-based procedure would be used. A content analysis approach was used to compare the evaluators' second evaluation comments to the experts' comments on strengths, areas for improvement, and site visit issues. The results of this content analysis were compared to the results of the content analysis done for Q5.

Question 11 asked "how might the training of evaluators be improved?" Answering the previous questions provided indications of relative variation and accuracy between categories. Testable hypotheses developed under Hypothesis 11 (H11) tested the relationship between subjects' accuracy and their perception of the difficulty of evaluating a category.

> H11: Subjects' perceived difficulty of evaluating a category will be negatively related to their accuracy in scoring that category.

The relationship proposed in H11 was tested for each category. Subjects' perceived difficulty of evaluating each category were collected by questionnaire during the second evaluations. A four point scale[21] was used to rate perceived difficulty: (1) easy,

[21] Four point scales, rather than five point scales, were used to force the respondent to make a decision favoring one extreme or the other.

(2) somewhat easy, (3) somewhat difficult, and (4) difficult. Tests were conducted using both DA and elevation as the measure of accuracy. An ordered categories method (Schulman, 1994) was used to conduct the tests. Hypothesis 12 (H12) addressed a related construct in similar fashion.

> H12: Subjects' perceived accuracy in scoring a category will be positively related to their accuracy in evaluating that category.

Testable hypotheses developed under H12 were tested for each category. Subjects' perceived accuracy in scoring each category were collected in the second evaluation questionnaire. To account for differing concepts of accuracy, subjects were asked to rate how close to the experts' scores they felt their scores were. A four point scale was used to rate perceived accuracy (i.e., closeness to expert scores): (1) remote, (2) somewhat remote,

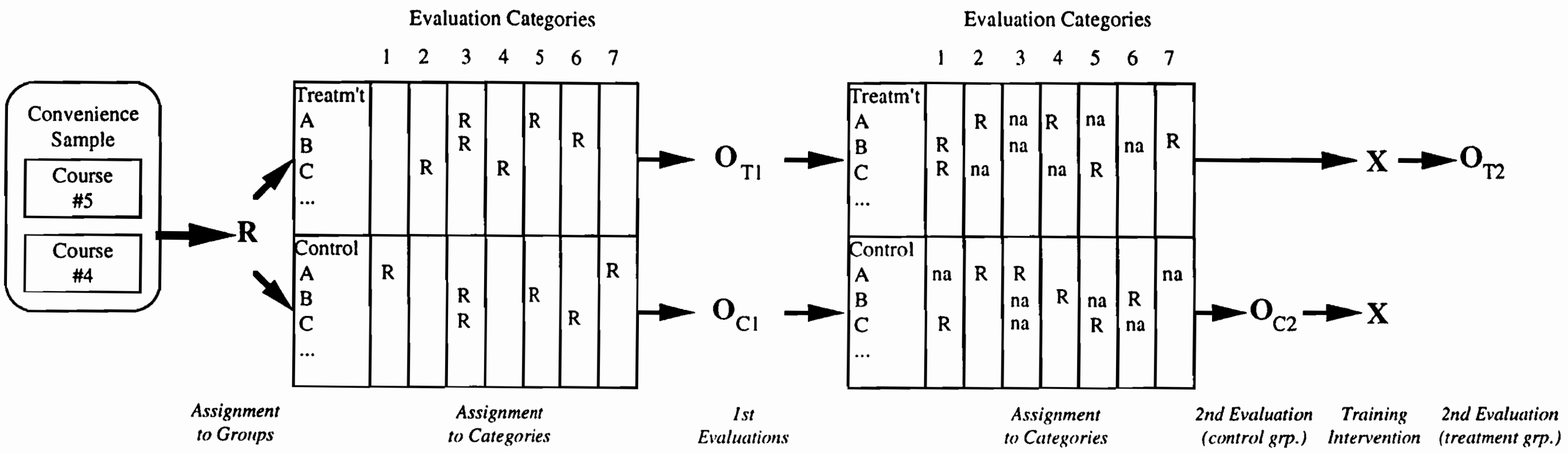
(3) somewhat close, and (4) close.

Question 12 asked "which of the following evaluator characteristics best predict the accuracy of evaluators' scores? previous use of the Baldrige (MB) criteria, level of education, educational specialty, amount of quality and productivity (Q&P) training, work experience, job function, job level, employer industry, employer size, age, and gender." Question 12 was not addressed by hypothesis testing, but was explored using multiple regression techniques. The data for each evaluator's characteristics was collected by questionnaire during the second evaluation. Results of the multiple regression were used to propose a model of the evaluator characteristics believed to affect the accuracy of evaluator scores. Descriptive statistics will be calculated for each characteristic, both by group and combined for all subjects. These descriptive statistics were used to further describe the subjects and to provide an indication of the range of characteristic values over which this model may be generalized.

Experimental Design

An experiment was designed to collect the data needed for the analyses described in the previous section. The experimental design is summarized in graphical form in Figure 2. A convenience sample of subjects were randomly assigned to either the treatment or control group. The subjects are described in the following section. Subjects in each group evaluated a case study on two randomly assigned evaluation categories. The subjects were given the criteria for evaluating these categories in advance and asked to familiarize themselves with the criteria and scoring system. Each subject was given the complete case study and a complete evaluation scorebook. Subjects were given one week to review the case study, evaluate the two categories, enter their evaluations into the scorebook, and submit their scorebooks to the researcher. Immediately afterwards the control group was given a second scorebook and asked to evaluate two different categories. The control group was also given a brief questionnaire to respond to after completing their second evaluation. The control group was given one week to evaluate the two additional categories, enter their evaluations into the scorebook, respond to the questionnaire, and submit their scorebooks and questionnaires to the researcher. The treatment, a training intervention, was then given to all subjects.

The training intervention was approximately two and one-half hours in length. The objectives and topics of the training intervention are given in the training plan (Appendix E). Following the treatment, the treatment group was given a second scorebook and asked to evaluate two different categories. The treatment group was also given the same questionnaire to respond to after completing their second evaluation. The treatment group was given one week to evaluate the two additional categories, enter their evaluations into the scorebook, respond to the questionnaire, and submit their scorebooks and questionnaires to the researcher. More detailed descriptions of the procedures and tools



Evaluation Categories

	1	2	3	4	5	6	7
Treatm't							
A			R		R		
B			R			R	
C		R		R			
...							
Control							
A	R						R
B			R		R		
C			R			R	
...							

Evaluation Categories

	1	2	3	4	5	6	7
Treatm't							
A		R	na	R	na		
B	R		na			na	R
C	R	na		na	R		
...							
Control							
A	na	R	R				na
B			na	R	na	R	
C	R		na		R	na	
...							

Notes:

(1) Due to the size of the case study, subjects only evaluated two categories per observation. Categories were randomly assigned to subjects.

(2) The first observation had subjects complete scorebooks for their assigned categories. The second observation had subjects complete scorebooks for their assigned categories and complete a survey of primarily demographic information.

(3) Because the treatment (training) was part of a course the subjects were taking, all subjects received the treatment. Therefore, the control group's second observation was taken immediately before and the treatment group's second observation immediately after the training.

(4) Key to symbols

R = a form of randomization procedure (random assignment of subjects to groups or random assignment of subjects to categories)
A, B, C = alphabetical listing of subjects assigned to that group
O = an observation or data collection point (submittal of scorebooks)
X = administration of the treatment (training intervention)
na = not available (i.e., subjects will not be able to evaluate the same category twice)

G. Coleman, 9/29/95; revised 6/27/96

Figure 2: Overview of the experimental design.

used for this experiment are given in the Procedures section and the accompanying appendices.

Subjects

The subjects for this experiment were 98 graduate students from two graduate level Industrial and Systems Engineering courses at a major research university in the southeastern United States. Seven of these students were registered in both courses[22]. Both courses were televised as part of an off-campus graduate program and will be referred to as course #4 and course #5. Quality and productivity management were major portions of the content in both courses. More than half of the subjects received the courses via television with two-way audio. While this provided a number of logistical challenges, it provided a sample dominated by full-time professionals.

Participation in the study was voluntary; however, the lecture on the Malcolm Baldrige National Quality Award and evaluation of the case study was part of the graded material of each course[23]. By agreeing to participate in the study, students were agreeing to allow their evaluations to be analyzed and to complete a questionnaire. A student's decision to participate or not participate in the experiment did not affect the grading of their case study evaluation. The potential grade advantage given the treatment group by providing training before their second evaluation was controlled by adjusting the grades of the control group (see Grading Procedures in Appendix F).

The demographic data collected via the questionnaire was used to construct a demographic profile of the subjects (see Chapters IV and V). While these subjects did represent a convenience sample, they were found to be representative of some strata of

[22] These seven students' responses were omitted from the analysis for the second of the two courses. More is said about screening of the data in Chapter IV.

[23] A memorandum of agreement was developed describing how the researcher and the instructors of these courses would collaborate on the guest lecture, case study evaluation, and research data collection (see Appendix R).

practicing quality and productivity improvement professionals. Nearly all of these subjects have professional work experience and responsibility relevant to the assessment and improvement of organizations.

Procedures

The tasks necessary to conduct this research were outlined in a work breakdown structure (WBS), shown in Appendix G. The remainder of this section elaborates on those tasks in the WBS requiring additional description. Tasks that are self-explanatory, clerical or administrative, or previously described receive minimal attention. The focus of the following paragraphs is on tasks that pertain to executing the steps of the experimental design. Tasks will be referred to by the coding scheme of the WBS. For example, writing this section of the research proposal is task I.E.2.c.3. Please note that the WBS includes five different levels of tasks and that the WBS does not portray a strict chronological sequence.

Task I, preparation, included the design, development, clerical, and administrative tasks necessary to execute data collection and analysis. The materials listed in task I.A. were obtained from NIST and ASQC in a timely manner. These materials were of very good quality, well suited for reproduction, and quite helpful for the development of the training materials. The MBNQA 1995 Award Criteria booklets contain the criteria and scoring system described in Chapter II. Appendix H displays the table of contents for the 1995 Award Criteria booklet. Each subject received a copy of this booklet at least two weeks prior to the beginning of data collection. The Colony Fasteners Case Study was prepared by NIST for use in the 1995 Malcolm Baldrige National Quality Award Examiner Preparation Course. The case study "is a sample application written for a fictitious company applying for the Baldrige Award" (NIST, 1995b). The case study is not a summary or mini-application, it uses the maximum allowed number of pages (seventy).

The case study represents an organizational self-assessment, prepared based on the criteria of the MBNQA. Each subject was given a complete copy of the Case Study. The table of contents of the Colony Fasteners Case Study is shown in Appendix I.

The sampling plan (I.B.) describes what data was collected from which subjects. All of the students in the two courses were asked to participate. Subjects will be entered into a single spreadsheet and sorted alphabetically. The detailed procedures used for assigning subjects to groups and assigning subjects to categories are described in Appendix S. These procedures identified which subjects would be evaluating which categories and when they would receive the training intervention. Two constraints were placed on the assignment process. No subject was assigned the same category twice, if the same number came up it was simply skipped. The number of subjects assigned each category was approximately balanced. A similar process was followed for assignment of categories for the second evaluation. The evaluator characteristics measured by the questionnaire (task I.B.4.) were discussed earlier.

The primary data collection forms were the scorebooks (I.C.1.) and questionnaires (I.C.2.). The scorebooks were a modified version of the MBNQA 1995 Application Scorebook[24], developed by NIST. The instructions were modified to remove references to returning the scorebook to the Award administrator, instructions regarding evaluating all seven categories, and steps related to consensus scoring (the second stage when using a team of examiners). The cover page was modified to refer to the first and second evaluation and to list the two categories assigned to the evaluator (see Appendix T for an example cover page). Each subject's name was pre-printed on the cover page of the scorebook along with a listing of the categories they were assigned.

The scorebook contains four types of worksheets and a brief description of how they are related. The Key Business Factors Worksheet (Appendix J) has the evaluator

[24] Example worksheets from the 1995 Application Scorebook are shown in Appendices J through M.

identify the areas of greatest relevance and importance to the applicant's business, assisting in the interpretation of the criteria as they apply to this particular organization. The Comment and Scoring Worksheet is the primary scoring worksheet. A separate Comment and Scoring Worksheet (see item 1.1 example, Appendix K) is provided for each item of the Award Criteria. Each Comment and Scoring Worksheet comprises two facing pages. The left page reiterates the description of the item[25], while the right page provides space to write comments on the applicant's strengths, areas for improvement, and site visit issues. The percent score given the applicant on this item is also written on the Comment and Scoring Worksheet. The Comment Summary Worksheet (Appendix L contains the modified version used for this study) asks the evaluator to summarize their overall evaluation of the applicant's response to the requirements of the 1995 Award Criteria. The Comment Summary Worksheets were not used for direct data collection. Subjects were instructed to keep these worksheets to assist in preparing their final summary comment memos[26]. The Score Summary Worksheet (Appendix M) contains space for listing all the scores for each item and category. Subjects were asked to list the scores for those items and categories they were assigned.

A questionnaire regarding evaluator characteristics and perceptions of the evaluation (Appendix N) was included with the scorebooks during the second evaluation. The questionnaires were attached to the cover memos accompanying the scorebooks. Subjects were not asked to list their names on the questionnaires. Each questionnaire was coded for identification purposes and distributed to the subjects along with the scorebooks which had their names on the cover page. The draft questionnaire was revised based on review by an expert on questionnaire construction and a pilot test with six test subjects.

[25] This description is a word-for-word reproduction of the item description from the 1995 MBNQA Criteria booklet.

[26] These final summary comment memos were part of the learning experience in each course, but were not a direct part of this study.

A request for approval for research involving human subjects was submitted to the Virginia Tech Internal Review Board and approved. The request asked for approval on an exempt (from full board review) basis. The informed consent statement developed for this request and used for this study is shown in Appendix O.

Task II included all data collection prior to the training intervention. This included the first and second evaluations for the control group and the first evaluation for the treatment group. An example of the announcement used for task II.A.1. is shown in Appendix P. Complete copies of the case study were distributed to each subject. A pre-assigned[27] application scorebook was given to each subject (II.B.1.). A cover memo was given to each subject asking them to familiarize themselves with the criteria and scoring system, read the case study, and then follow the instructions in the scorebook to evaluate their assigned categories. Appendix U contains an example cover memo. All subjects were given one week to complete this first evaluation. A complete data collection and training schedule for both courses, covering elements of tasks II, III, and IV, is shown in Appendix Q. Actual execution closely followed this schedule, with only minor deviance from the shipping and receiving dates. The evaluations were graded per the grading plan in Appendix F. Expert evaluation of the case study including key business factors, percent scores for each item, strengths, areas for improvement, and site visit issues were taken from the Colony Fasteners Evaluation Notes prepared by NIST.

A similar distribution procedure was followed for the control group's second evaluation (II.E.) in the week immediately following the first evaluation. The second evaluation included the administration of the questionnaire described earlier (see I.C.2.). In addition to the instructions for completing the second evaluation and questionnaire, the cover memo reminded the subjects that approximately half the class would be doing their

[27] Each scorebook will be pre-printed with the subject's name and assigned categories on the cover. See Appendix T for an example.

second evaluation now and the other half following the lecture(s) on the Malcolm Baldrige National Quality Award and its uses. The memo also reiterated the use of a normalized grading procedure that levels the playing field for those assigned to either group. At the same time, the treatment group received a memo reminding them that they would be conducting their second evaluation in the near future and reiterating the normalized grading procedure. Each memo was individually addressed to ensure proper distribution. Appendices V and W contain examples of these memos. The control group was given one week to complete their second evaluation and submit their scorebooks and questionnaires.

Task III was the training intervention (experimental treatment). The student objectives and topical outline (III. A.) for the training are shown in Appendix E. The training intervention included lecture and in-class exercises. The lecture was based on the 1994 and 1995 MBNQA materials from NIST (including the Handbook for Examiners), Reimann's (1992) and Wachniak's (1990) historical perspectives, Bemowski and Stratton's 1995 survey, the First European Forum on Quality Self-Assessment (1994), the self-regulation model presented in Chapter I, and the Great Northern Case Study (NIST, 1994c). Excerpts from the Great Northern Case Study were used for exercises to practice evaluating and scoring using the MBNQA criteria and scoring system[28]. The class was presented an overview of the Case Study and the key business factors for Great Northern. Vignettes based on selected items[29] were presented one at a time and the subjects were asked to individually evaluate and score the vignettes. Next, the class shared and discussed their scores for that item. Then the experts' scores and comments for the item were presented and discussed[30]. Since Great Northern is based on the 1994 criteria, it was

[28] These exercises represent frame-of-reference (FOR) training, as used in rater research.

[29] Two items from the Great Northern Case Study were selected: Item 1.1 - Senior Executive Leadership and Item 6.3 - Supplier Performance Results. These were chosen because they use the two different scoring guidelines (approach/deployment for Item 1.1 and results for Item 6.3) and because they represented extremes in scoring (the experts scored Item 1.1 at 75% and Item 6.3 at 35%).

[30] Reaction to this portion of the training was quite enthusiastic. Since this was an ungraded in-class exercise, the students seemed to approach it like a game of trying to "match the experts." The interaction

desirable to select items for the vignettes that had changed little between 1994 and 1995. It was also desirable to select items that were representative of the range of scores and items that used different scoring guidelines. It was not possible to meet both these desires; therefore, adjustments were made[31] and highlighted during presentation of the expert evaluation. An example of one of the items used (Item 6.3) is shown in Appendix X. Note that the 1994 criteria for Item 6.3 (and therefore, the Great Northern Case Study) included two areas to address, but these were combined into a single area to address for the 1995 criteria. The text of Great Northern's Item 6.3 response was simply reorganized to correspond with this single area to address. The lecture summary included a synopsis of this research project and its design.

Task IV, post-training data collection, included the second evaluation and questionnaire for the treatment group and the final summary comment memos from all subjects. The cover memo for the treatment group was similar to that provided the control group for their second evaluation, with the addition of reiterating the assignment of the final summary comment memo (see grading plan in Appendix F for description). The control group received a memo thanking them for their participation and reminding of the assignment of the final summary comment memo (see Appendix Y for an example of this memo). The treatment group was given one week[32] to complete their second evaluation and submit their scorebooks and questionnaires. Both groups were given one week to submit their final summary comment memos.

and degree of participation from those at remote sites appeared to be higher than that usually observed in televised courses.

[31] The primary changes in the item criteria were related to the organization of the areas to address, not the content of the criteria. The adjustments consisted of reorganizing the content of the case study within the items to reflect the organization of the areas to address in the 1995 criteria.

[32] For Course #5, the Thanksgiving Holiday occurred during the treatment group's second evaluation, forcing the due date to be extended. This resulted in an elapsed time of twelve days from distribution to submittal; however, the subjects were instructed to spend no more than seven days on the evaluation. Examination of their responses provided no evidence that they spent any more time on the evaluation than the treatment group from Course #4.

Task V, data analysis, began with data entry (tasks V.A. and V.B.). The analyses were conducted as described in the research hypotheses section earlier in this chapter and described in detail in Chapter IV. The software packages used as tools for these analyses are listed in Appendix Z. Analyses were conducted and results interpreted on a question-by-question basis, including supporting research and testable hypotheses. Results of the analyses were compiled and preliminary conclusions identified. Final conclusions were drawn based on the gestalt of the results. The communication of these results and conclusions is described by task VI. The experiment and this report ended with conclusions and recommendations for future research.

Chapter IV. Data Analyses

This chapter presents the analysis of the data on a question by question basis. The raw data and intermediate outputs are contained in the appendices (see Appendices AA to AU and Appendix BA). Summary statistics and graphical presentations of the data are contained in the tables and figures accompanying each research question.

Question 0 - For a given sample of untrained evaluators, what is the consistency of their scores?

This broad question examined the relative location of evaluator scores between groups, the dispersion of item and category scores between groups, the dispersion of scores across items in a category, and the dispersion of scores across categories. This question was addressed by comparing the evaluators' first evaluation (pre-treatment) scores on each item and category, testing hypothesis one (H1), and addressing research questions one through four. The evaluators' scores on each item and each category can be compared by examining the box plots of the scores (Figure 3) and the descriptive statistics in Appendix AA. For a given item or category, each box plot shows the distribution of the pre-training scores of the control group beside the distribution of the pre-training scores of the treatment group. The analyses of H1 and research question one are covered in this section, the other research questions are contained in the succeeding sections.

H1: There will be no difference in scores between the treatment and control groups during the initial evaluation: a) by item and b) by category

Hypothesis one addressed the first set of research questions related to consistency, rather than addressing a specific research question. H1 was intended to establish the baseline condition, that there were no differences between the mean scores of the treatment and control groups prior to the administration of the treatment. H1 was addressed by testing the specific testable hypotheses shown below. H1 was tested at both the item and category levels.

Testing Hypothesis 1

T-tests were used to compare the first evaluation mean scores of the control and treatment groups on each item and category[1]. Prior to analysis, the data were screened to remove responses suspected of contamination. Since this was a comparison of first evaluation scores, possible contamination of a subject's second evaluation scores was not considered relevant. Although Minitab was used to calculate actual p-values, a test-wise Type I error rate of alpha = 0.05 was used to identify significant results. Under the null hypotheses, it was reasonable to expect the variance of the scores to be the same; therefore, a pooled variance estimate was used.

Data Screening

Those subjects participating in both courses were identified and all their responses from the second course were discarded. Their evaluations from the second course were likely affected by their experience in the earlier course. Their second evaluations from the first course were also discarded. The earliest submission date for a second evaluation in the first course was two days after distribution of the first evaluation materials in the second course. This overlap may have provided additional experience to those in both courses; therefore, their second evaluations in the first course may have been contaminated. Since paired comparisons were not being performed, these second evaluations were independent and should have no impact on the validity of the scores from their first evaluations. The screened data (item scores) for the control and treatment groups' first evaluation are shown in Appendices BA and BB, respectively.

[1] Category scores were calculated based on the average of the category's item scores.

Tested Hypotheses

The following hypothesis was tested for each item.

H_0: There is no difference between the Item X.X mean scores from two randomly split-halves of a sample of untrained evaluators.
H_1: There is a difference between the Item X.X mean scores from two randomly split-halves of a sample of untrained evaluators.
where X.X = 1.1 to 7.5

The following hypothesis was tested for each category.

H_0: There is no difference between the Category X.0 mean scores from two randomly split-halves of a sample of untrained evaluators.
H_1: There is a difference between the Category X.0 mean scores from two randomly split-halves of a sample of untrained evaluators.
where X.0 = 1.0 to 7.0 and each evaluator's Category score is the average of their Item scores.

Test Results

Table 2 provides a summary of the t-test results by item and category. Descriptive statistics and detailed results of the t-tests are shown in Appendix AA. Detailed t-test results are provided for each item in a category, followed by the t-test for that category's average scores, followed by the descriptive statistics for that category. Box plot comparisons of the control and treatment groups' pre-treatment scores by item and category are provided in Figure 3. Each page of Figure 3 contains the box plot comparisons for all the items of a category and the box plot comparison of the average scores for that category.

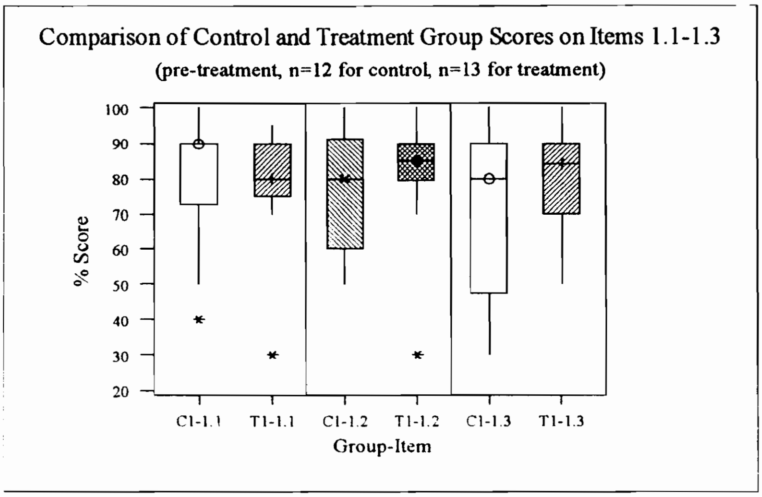
Table 2.

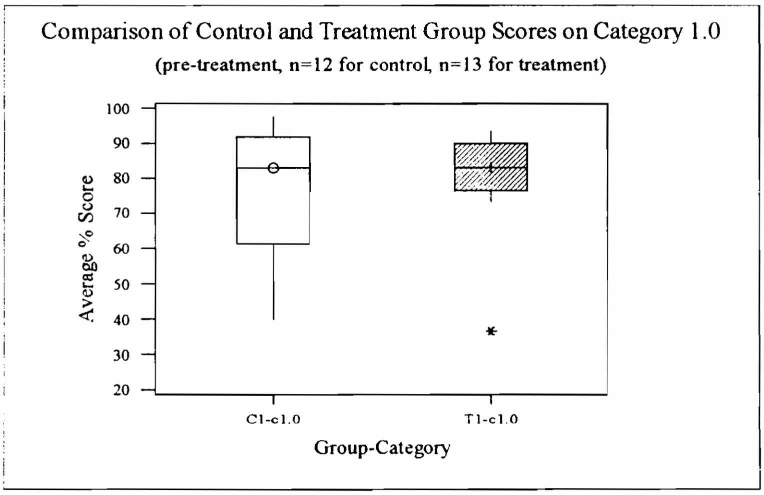
Summary of Test Results for Hypothesis 1

t-test for	P-Value (all data)	P-Value (screened)	Sample Size (screened) (control, treatment)
- Item 1.1	0.97	0.83	12, 13
- Item 1.2	0.65	0.66	12, 13
- Item 1.3	0.47	0.29	12, 13
- Item 2.1	0.55	0.80	12, 11
- Item 2.2	0.29	0.38	12, 11
- Item 2.3	0.091	0.17	12, 11
- Item 3.1	0.46	0.42	11, 13
- Item 3.2	0.36	0.51	11, 13
- Item 4.1	0.88	1.00	11, 14
- Item 4.2	0.84	0.72	11, 14
- Item 4.3	0.20	0.24	11, 14
- Item 4.4	0.64	0.58	11, 14
- Item 5.1	0.58	0.71	9, 11
- Item 5.2	0.12	0.20	9, 11
- Item 5.3	0.40	0.51	9, 11
- Item 5.4	0.20	0.27	9, 11
- Item 6.1	0.95	0.95	7, 12
- Item 6.2	0.50	0.50	7, 12
- Item 6.3	0.77	0.77	7, 12
- Item 7.1	0.31	0.31	12, 12
- Item 7.2	0.40	0.40	12, 12
- Item 7.3	0.30	0.30	12, 12
- Item 7.4	0.14	0.14	12, 12
- Item 7.5	0.085	0.085	12, 12
- Category 1.0	0.61	0.61	12, 13
- Category 2.0	0.15	0.27	12, 11
- Category 3.0	0.35	0.42	11, 13
- Category 4.0	0.56	0.56	11, 14
- Category 5.0	0.21	0.31	9, 11
- Category 6.0	0.65	0.65	7, 12
- Category 7.0	0.084	0.084	12, 12

The initial p-values were calculated using all data collected and were provided for comparison purposes only. The second p-values were calculated following screening of data. The second p-values were used for experimental interpretation.

Category 1.0 - Leadership





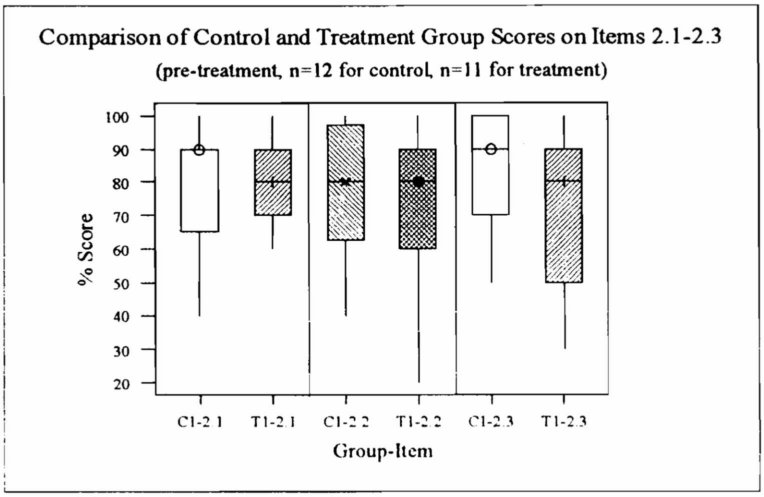
Explanation of coding for the x-axis.

The data sets for these and subsequent box plots are labeled by Group-Item or Group-Category. Thus, "C1-1.1" refers to the data from the control group's first evaluation of item 1.1. "T1-1.2" refers to the data from the treatment group's first evaluation of item 1.2. "C1-c1.0" refers to the data (average of the item scores) from the control group's first evaluation of category 1.0. Subsequent figures will include data from the second evaluation (e.g., "C2-1.1").

Figure 3. Box plot comparisons of control and treatment groups' pre-treatment scores.

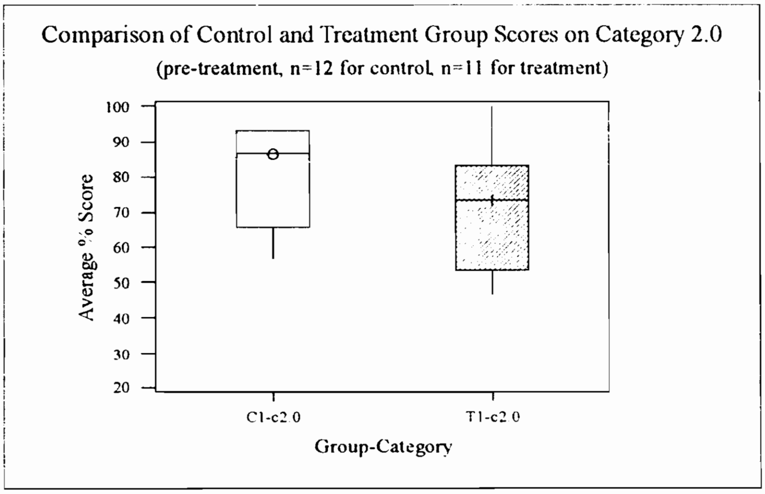
figure continues

Category 2.0 - Information and Analysis



Comparison of Control and Treatment Group Scores on Items 2.1-2.3

(pre-treatment, n=12 for control, n=11 for treatment)

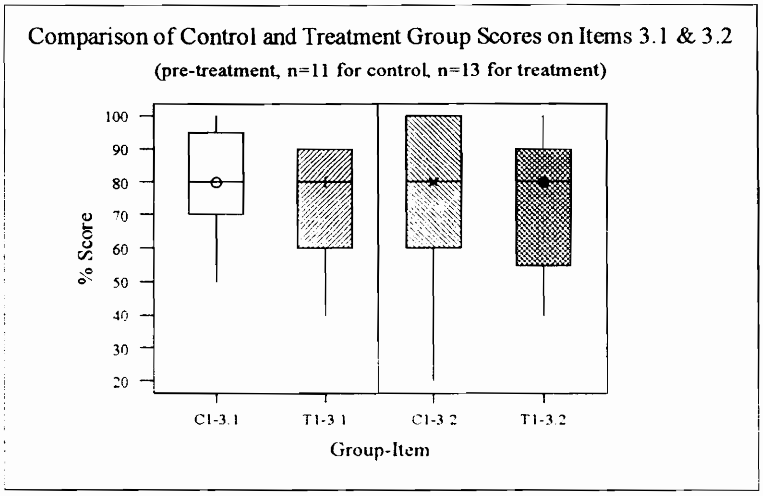


Comparison of Control and Treatment Group Scores on Category 2.0

(pre-treatment, n=12 for control, n=11 for treatment)

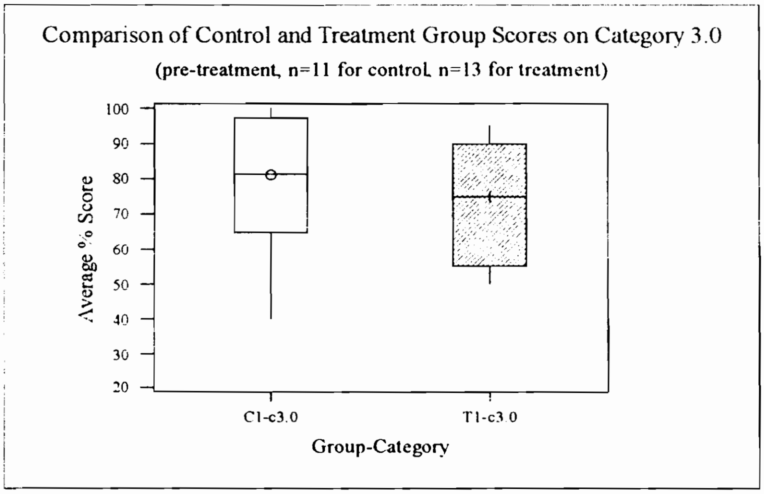
figure continues

Category 3.0 - Strategic Planning



Comparison of Control and Treatment Group Scores on Items 3.1 & 3.2

(pre-treatment, n=11 for control, n=13 for treatment)

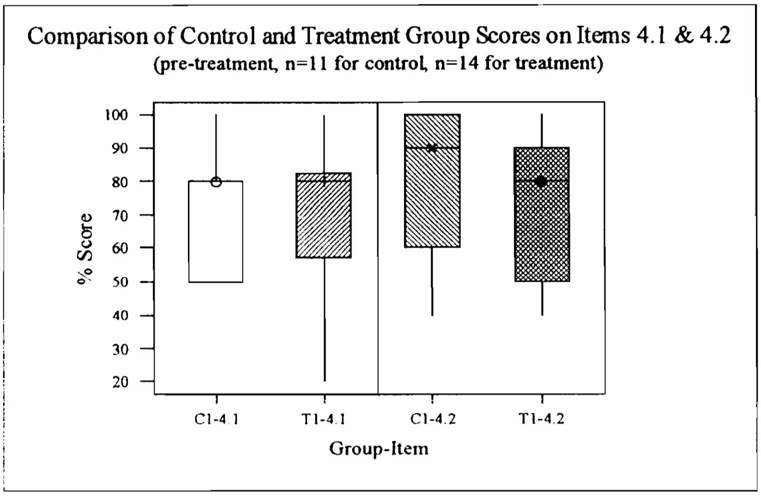


Comparison of Control and Treatment Group Scores on Category 3.0

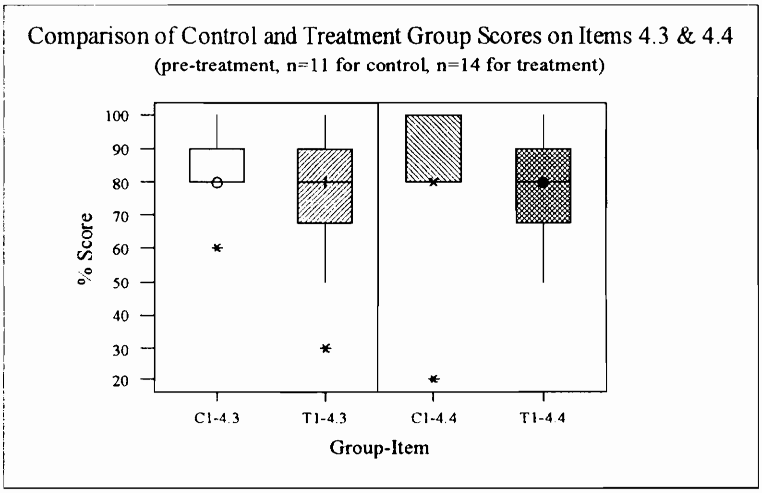
(pre-treatment, n=11 for control, n=13 for treatment)

figure continues

Category 4.0 - Human Resource Development and Management

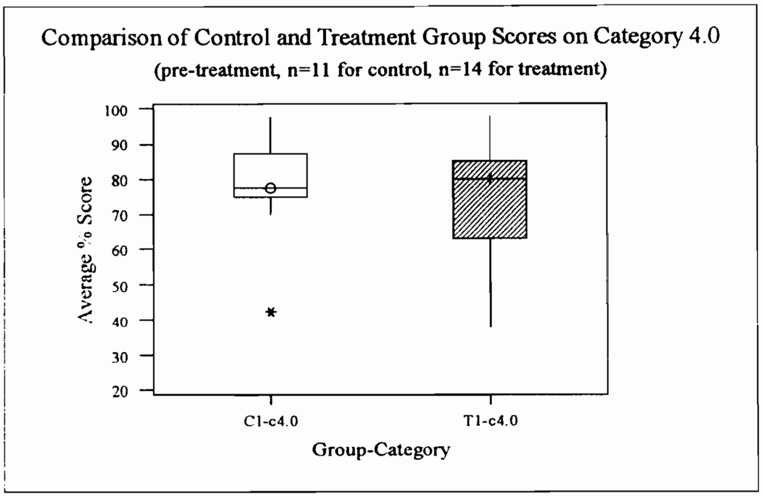


Comparison of Control and Treatment Group Scores on Items 4.1 & 4.2
(pre-treatment, n=11 for control, n=14 for treatment)



Comparison of Control and Treatment Group Scores on Items 4.3 & 4.4
(pre-treatment, n=11 for control, n=14 for treatment)

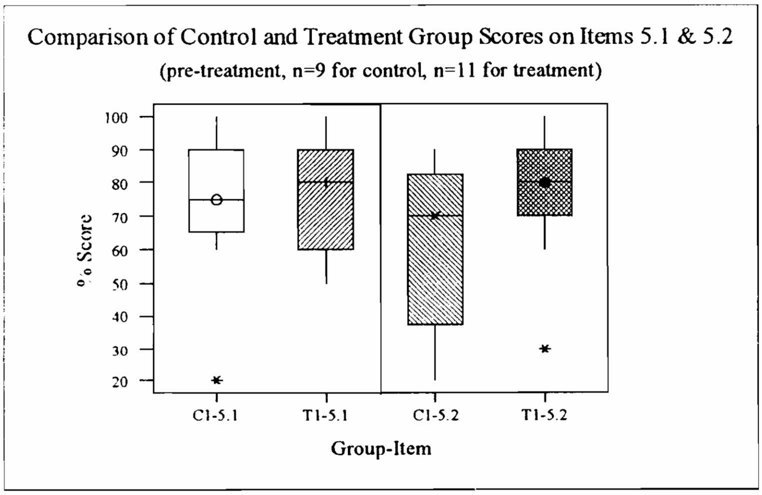
figure continues



Comparison of Control and Treatment Group Scores on Category 4.0

(pre-treatment, n=11 for control, n=14 for treatment)

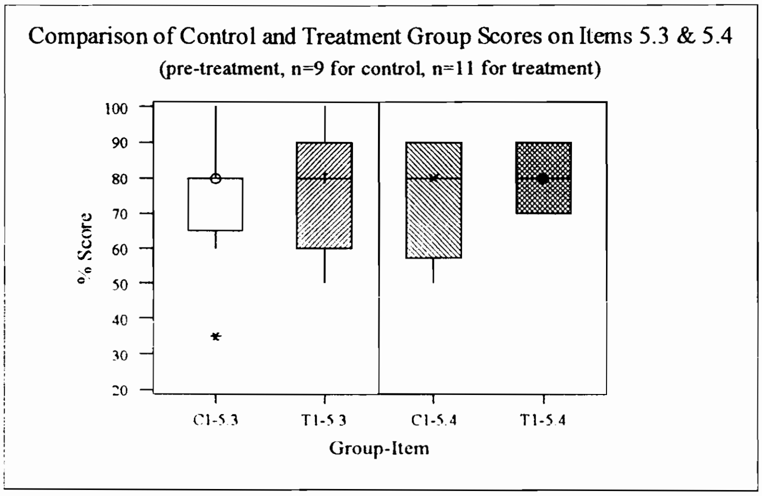
Category 5.0 - Process Management



Comparison of Control and Treatment Group Scores on Items 5.1 & 5.2

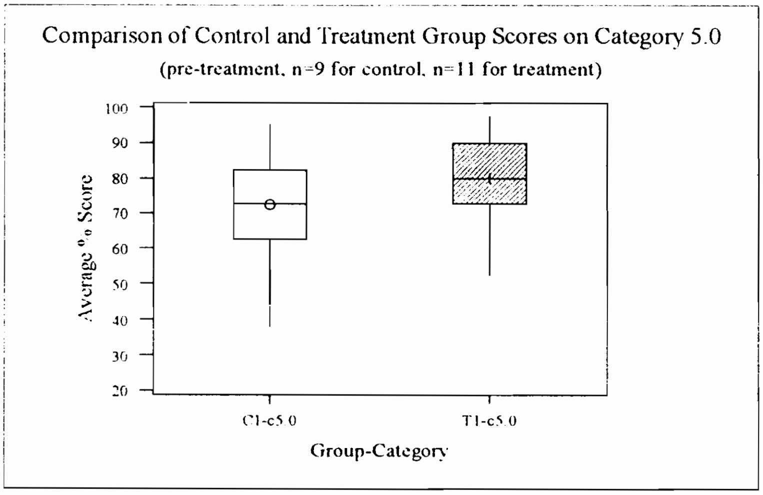
(pre-treatment, n=9 for control, n=11 for treatment)

figure continues



Comparison of Control and Treatment Group Scores on Items 5.3 & 5.4

(pre-treatment, n=9 for control, n=11 for treatment)

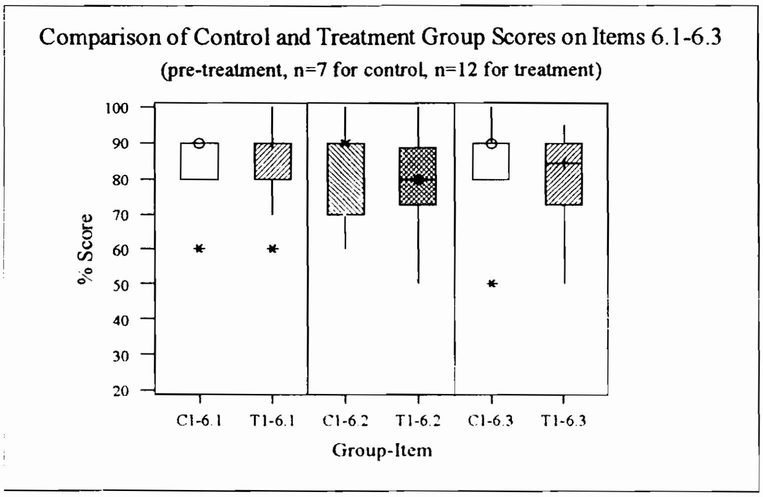


Comparison of Control and Treatment Group Scores on Category 5.0

(pre-treatment, n=9 for control, n=11 for treatment)

figure continues

Category 6.0 - Business Results


Comparison of Control and Treatment Group Scores on Items 6.1-6.3
(pre-treatment, n=7 for control, n=12 for treatment)

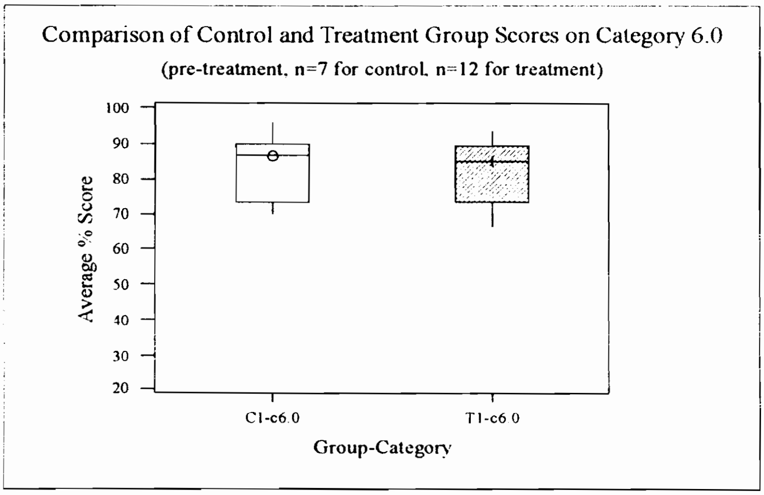
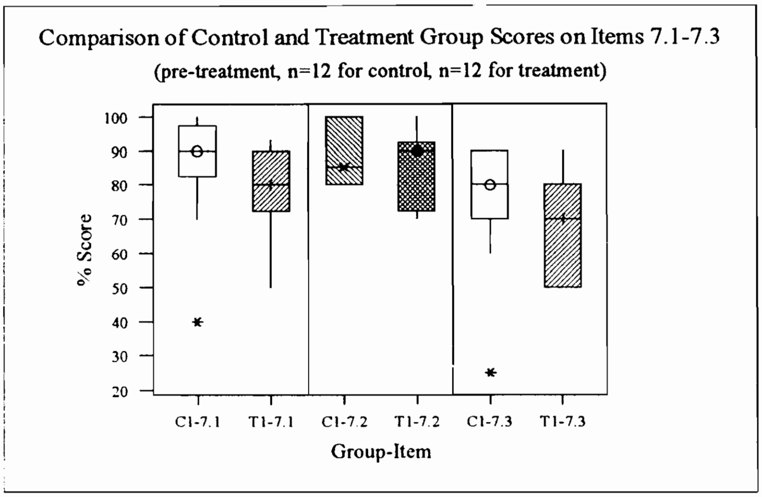

Comparison of Control and Treatment Group Scores on Category 6.0
(pre-treatment, n=7 for control, n=12 for treatment)

figure continues

Category 7.0 - Customer Focus and Satisfaction



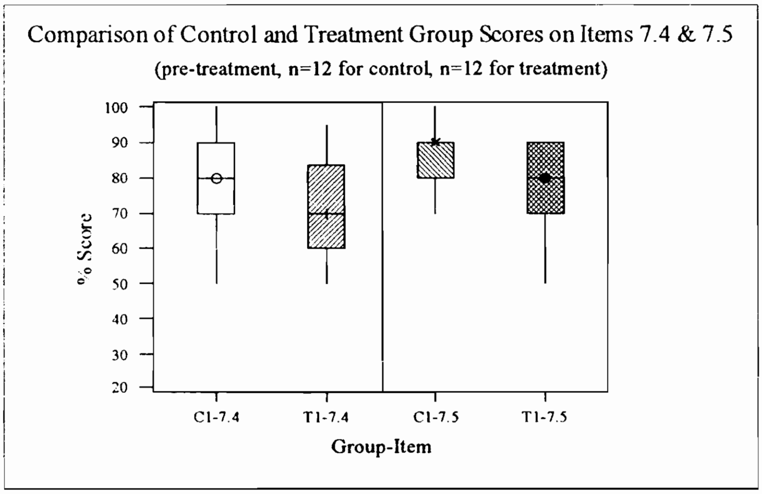
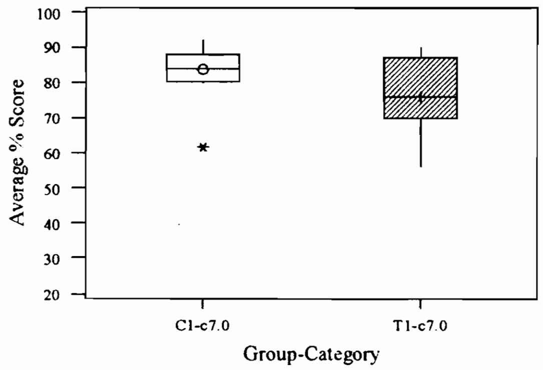


figure continues

Comparison of Control and Treatment Group Scores on Category 7.0

(pre-treatment, n=12 for control, n=12 for treatment)



Question 1 - How much agreement is there among evaluators on the score of an item?

Question 1 (Q1) examined the relative agreement (i.e., lack of dispersion of scores) within and between the control and treatment groups on the score of each item. Q1 was addressed by examining the relative agreement displayed by the evaluators on the score of each item and by testing hypothesis 2 (H2). The relative agreement of each group's untrained evaluators on the score of each item can be seen in the box plots of the item scores (Figure 3). Hypothesis 2 tested for differences in item score variance between the control and treatment group. Since evaluators were presumed equal prior to administration of the treatment, no difference was expected to be seen between the score variances of the control and treatment groups.

H2: For each item, there will be no difference in score variances between the treatment and control groups.

H2 was tested using the treatment and control groups' scores prior to the administration of the treatment (i.e., first evaluation scores). The specific hypotheses tested to address H2 are shown below.

Testing Hypothesis 2

F-tests were used to compare the first evaluation score variances of the control and treatment groups on each item. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 1). A test-wise Type I error rate of alpha = 0.05 was used.

Tested Hypotheses

The following hypothesis was tested for each item.

H_0: There is no difference between the variances of Item X.X scores from two randomly split-halves of a sample of untrained evaluators.
H_1: There is a difference between the variances of Item X.X scores from two randomly split-halves of a sample of untrained evaluators.
where X.X = 1.1 to 7.5

Test Results

Table 3 provides a summary of the F-test results by item. Descriptive statistics including interquartile range are shown in Appendix AA. The interquartile range of each group's scores on each item are also shown in the box plot comparisons of
Figure 3.

Table 3.

Summary of Test Results for Hypothesis 2.

test of variances for Item	sample size (C1, T1)	p-value
1.1	12, 13	0.6854
1.2	12, 13	0.9286
1.3	12, 13	0.0670
2.1	12, 11	0.1274
2.2	12, 11	0.3402
2.3	12, 11	0.2676
3.1	11, 13	0.6350
3.2	11, 13	0.4184
4.1	11, 14	0.5268
4.2	11, 14	0.6738
4.3	11, 14	0.0502
4.4	11, 14	0.2444
5.1	9, 11	0.3830
5.2	9, 11	0.4058
5.3	9, 11	0.7716
5.4	9, 11	0.0560
6.1	7, 12	0.8804
6.2	7, 12	0.7234
6.3	7, 12	0.6908
7.1	12, 12	0.4664
7.2	12, 12	0.7178
7.3	12, 12	0.4988
7.4	12, 12	0.9494
7.5	12, 12	0.3976

Question 2 - How Much Agreement is There Among Evaluators on the Score of a Category?

Question 2 (Q2) was addressed by examining the relative agreement (i.e., lack of dispersion of scores) within the control and treatment groups on the score of each category, by comparing the relative agreement displayed by the two groups on the score of each category, and by testing hypothesis 3 (H3). The relative agreement of the untrained evaluators on the score of each category can be seen in the box plots of the category scores (Figure 3). Hypothesis 3 tested for differences in category score variance between the control and treatment group. Since evaluators were presumed equal prior to administration of the treatment, no difference was expected to be seen between the score variances of the control and treatment groups.

H3: For each category, there will be no difference in score variances between the treatment and control groups.

H3 was tested using the treatment and control groups' scores prior to the administration of the treatment (i.e., first evaluation scores). The specific hypotheses tested to address H3 are shown below.

Testing Hypothesis 3

F-tests were used to compare the first evaluation score variances of the control and treatment groups on each category. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 1). A test-wise Type I error rate of alpha = 0.05 was used.

Tested Hypotheses

H_0: There is no difference between the variances of Category X.0 scores from the split-halves of a sample of untrained evaluators.

H_1: There is a difference between the variances of Category X.0 scores from the split-halves of a sample of untrained evaluators.
where X.0 = 1.0 to 7.0

Test Results

Table 4 provides a summary of the F-test results by category. Descriptive statistics including interquartile range are shown in Appendix AA. The interquartile ranges of the control and treatment groups' pre-treatment scores by category were provided in the box plots of Figure 3.

Table 4.

Summary of Test Results for Hypothesis 3.

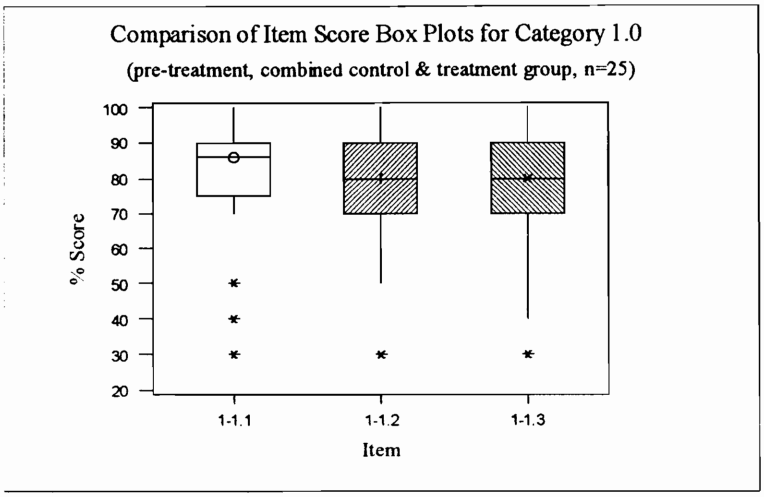
test of variances for Category	sample size (C1, T1)	p-value
1.0	12, 13	0.4626
2.0	12, 11	0.6710
3.0	11, 13	0.7540
4.0	11, 14	0.5410
5.0	9, 11	0.5852
6.0	7, 12	0.8704
7.0	12, 12	0.2760

Question 3 - How consistent is the within-item variation of the evaluator scores across all the items of a category?

Question 3 (Q3) examined the consistency of score dispersion between the items of a category and was examined for each category. Q3 was addressed by comparing the relative variation of item scores across all the items of a category. Because the variance of item scores within a category were not independent, a testable hypothesis was not developed to address Q3. Since no statistical differences were found in the (pre-treatment) item score variances between the control and treatment group, the group scores were pooled for this analysis[2]. The scores of the control and treatment group were pooled for each item and then used to produce box plots. The box plots for all the items of a category were plotted on a single chart to facilitate comparison (see Figure 4). The data used for these box plots were the screened data from H1. Since the items within a category measure related constructs, little or no difference was expected between the dispersions of item scores across a particular category. This effect (or lack thereof) may have been reinforced by the limited ability of untrained evaluators to discriminate among related constructs.

[2] Since subjects were randomly assigned to groups, an argument could be made that the pre-treatment scores of the two groups could be pooled even if statistical differences had been found.

Category 1.0 - Leadership



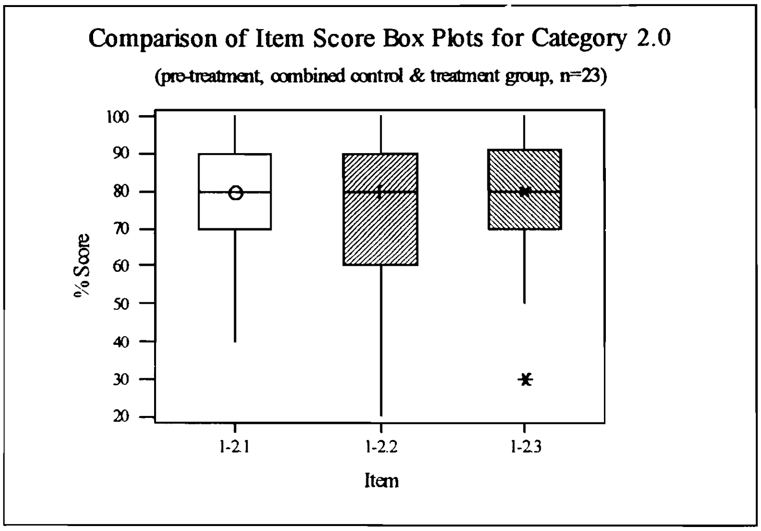
Explanation of coding for the x-axis.

The data sets for past box plots were labeled by Group-Item or Group-Category. Thus, "C1-1.1" referred to the data from the control group's first evaluation of item 1.1. Since these box plots were produced with combined data from the control and treatment group, labels refer to the evaluation and item. Thus, "1-1.1" refers to the combined data from the first evaluation of item 1.1.

Figure 4. Box Plot Comparisons of Item Scores Across Each Category (pre-treatment).

figure continues

Category 2.0 - Information and Analysis



Category 3.0 - Strategic Planning

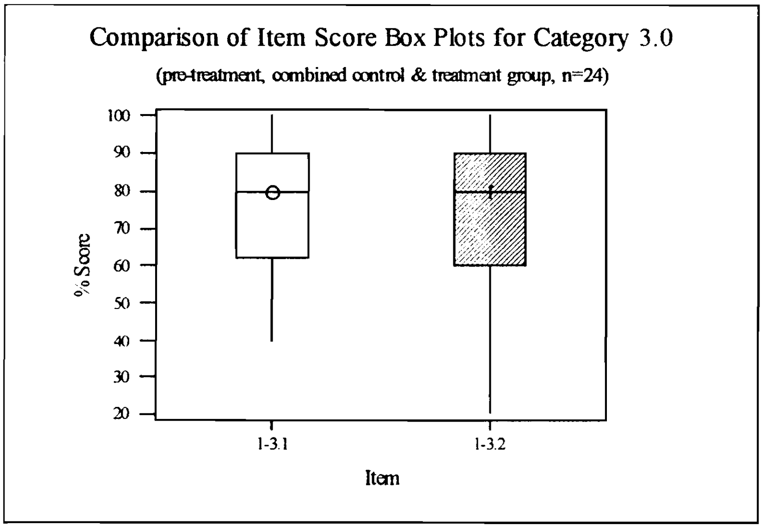
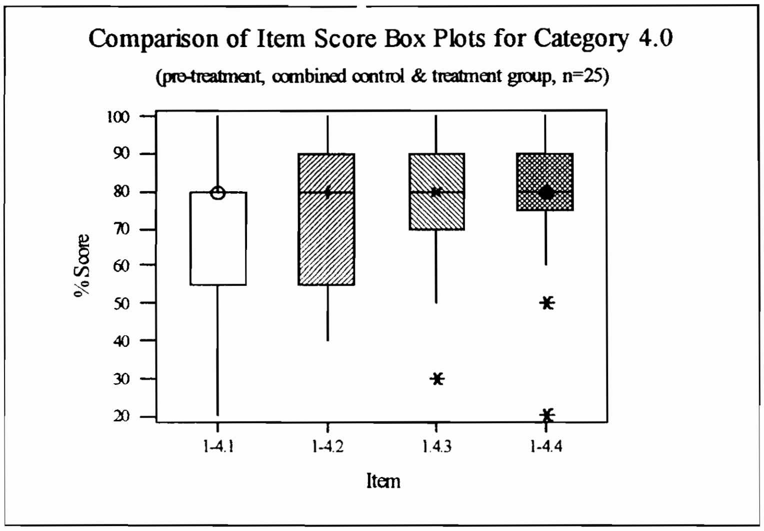


figure continues

Category 4.0 - Human Resource Development and Management



Category 5.0 - Process Management

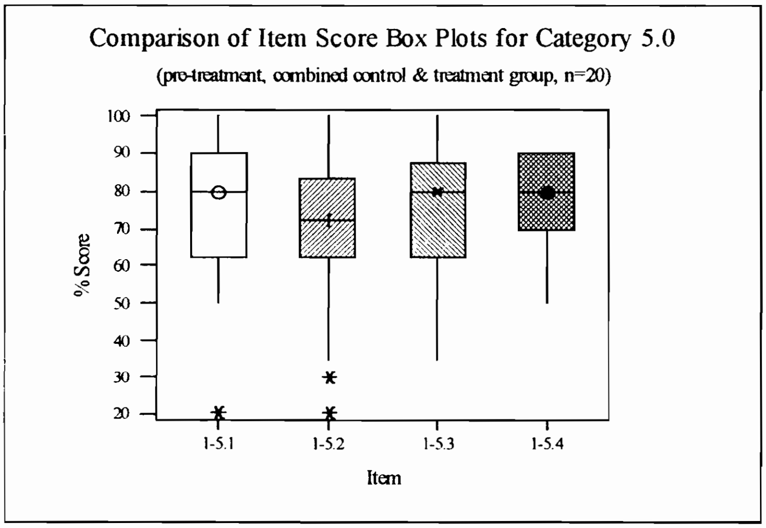
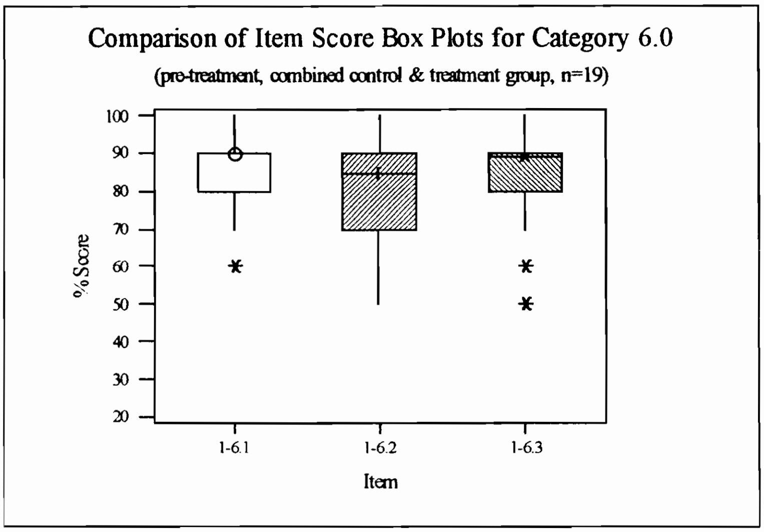
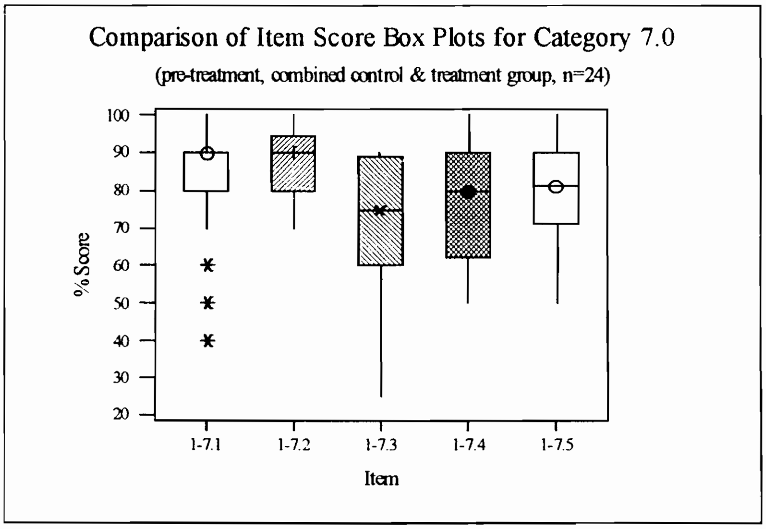


figure continues

Category 6.0 - Business Results



Category 7.0 - Customer Focus and Satisfaction



Question 4 - How consistent is the within-category variation of the evaluator scores across all seven categories?

Question 4 (Q4) examined the consistency of category score dispersion across all seven categories. Q4 was addressed by testing hypothesis 4 (H4) and graphically comparing the relative dispersion of category scores across all seven categories. Hypothesis 4 tested for differences in category score variances between the seven categories. Since each category measures a different construct, a difference was expected to be seen between the score variances of the seven categories.

H4: There will be a difference in score variances between categories.

H4 was tested using the pooled control and treatment groups' pre-treatment scores from Q3. The specific hypothesis tested to address H4 is shown below. The relative dispersion of category scores across all seven categories can be seen in the box plots of Figure 5.

Testing Hypothesis 4

Hartley's F_{max} test (Ott, 1984) was used to test for homogeneity of category variances. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 1). A Type I error rate of alpha = 0.05 was used. Since the overall test for differences in score variances between categories was inconclusive (i.e., borderline significant), pairwise comparisons were conducted to provide additional evidence or lack of evidence of differences in category score variances.

Tested Hypothesis

The following hypothesis was tested using the combined first evaluation score data.

H_0: There is no difference in the variances of category scores across the seven categories.
H_1: There is a difference in the variances of category scores across the seven categories.

The test statistic and rejection region for H4 are described in Appendix AV.

Test Results

Table 5 lists the standard deviations and resulting variances of the scores for each category. The standard deviation of the scores for each category were taken from Appendix AB: Edited Minitab Session Files for Hypothesis 4. The maximum and minimum variances from Table 5 were used to produce the F_{max} statistic below.

$F_{max\ obs.} = 3.69$

Comparing the observed statistic ($F_{max\ obs.}$) to the closest available critical value ($F_{max(7,\ 20)0.95}$) with less than or equal to the actual degrees of freedom implies a non-significant result; however, $F_{max\ obs.}$ falls between two $F_{max\ critical}$ values with close to the actual degrees of freedom (see Appendix AV). While linear interpolation of $F_{max\ critical}$ values is of questionable validity, it may give an indication of whether further testing is necessary. Linear interpolation between $F_{max(7,\ 20)0.95}$ and $F_{max\ (7,\ 30)0.95}$ yields an estimate of $F_{max(7,\ 23)0.95}$ equal to 3.66. If this is a valid estimate, the test would indicate significant results and H_0 would be rejected. These conflicting interpretations of the test result appear to justify conducting pairwise comparisons of the categories' score variances.

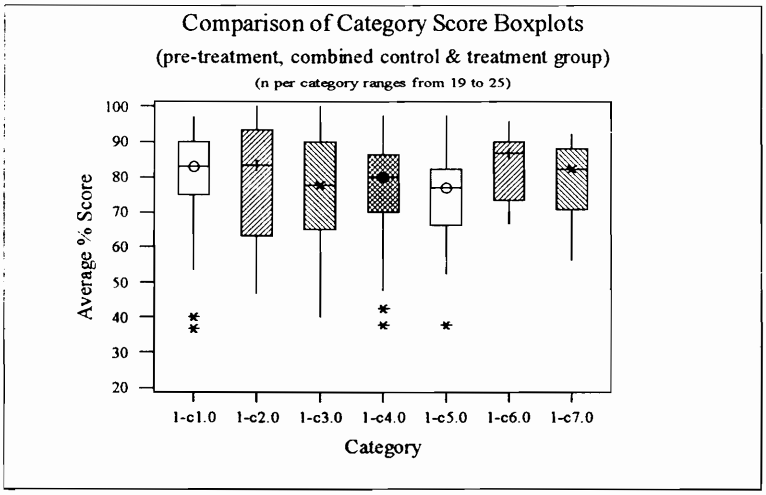


Figure 5. Box plot comparisons of category scores (pre-treatment).

Table 5.

Standard Deviations and Variances of the Scores of each Category (first evaluation)

Category	Std. Dev.	Variance	F-obs	n
1-c1.0	16.23	263.41		25
1-c2.0	15.95	254.40		23
1-c3.0	17.42	303.46	numerator	24
1-c4.0	15.94	254.08		25
1-c5.0	14.96	223.80		20
1-c6.0	9.07	82.26	denominator	19
1-c7.0	9.82	96.43		24
			3.69	
Note: These standard deviations are from the combined control and				
treatment group scores. All post-screening data points were used.				

Pairwise Comparisons (H4)

Pairwise comparisons were conducted to determine which, if any, category score variances were different. The variance of the scores of each category was compared to the variance of the scores of each of the other six categories. The data used for these comparisons were the screened data from Hypothesis 1, further screened to reduce possible dependencies. Since each subject evaluated two categories, using scores from the same subject for both categories of a comparison violates the assumption of independence. To compensate for this, the data were screened as described below.

Data Screening to Eliminate Dependencies

The following procedure was used for each paired comparison. The data for all the subjects who evaluated either of the two categories being compared were compiled into a spreadsheet. Those subjects evaluating both categories were numbered 1 to N. The next step was to assign each of these subjects to one of two groups, each group N/2 in size. Microsoft Excel's random number generator was used to produce a table of random numbers with a uniform distribution and a range greater than or equal to 1 to N. A new column of random numbers was used for each comparison, starting alternately at the top or bottom of the column. Each random number was used to identify which of the 1 to N subjects were assigned to the first group (i.e., their scores from the second category would be discarded for this analysis). Those remaining after N/2 had been assigned to the first group were assigned to the second group (i.e., their scores from the first category would be discarded for this analysis). When N was an odd number, a coin flip was used to determine which group would receive the additional subject. The result was elimination of the dependency while maintaining sample size as much as possible. Subsequent comparisons started with the complete data set and followed the same screening procedure.

Tested Hypotheses

F-tests were used to compare the score variances of each pair of categories. An experiment-wise Type I error rate of alpha = 0.05 was used. With the Bonferroni adjustment, this resulted in a comparison-wise error rate of alpha = 0.0024. The following hypothesis was used to compare each pair of categories.

H_0: There is no difference between the variance of Category X.0 scores and the variance of Category Y.0 scores.
H_1: There is a difference between the variance of Category X.0 scores and the variance of Category Y.0 scores.
where X.0 = 1.0 to 6.0 and Y.0 = 2.0 to 7.0 (no category was compared against itself).

Test Results

Table 6 lists the statistics and results for each comparison. Since the comparison-wise error rate was relatively small, those p-values less than the experiment-wise Type I error rate were highlighted for informational purposes. Table 7 provides a summary of the p-values in matrix form to facilitate evaluation. That is, Table 7 provides the results in a form that highlights patterns across the pairwise comparisons. See Appendix AC for the master spreadsheet and an example spreadsheet used to document the screening and calculate the variances for the pairwise comparisons.

Table 6.

Test Results for H4 Pairwise Comparisons

Comparison of variances for Category X.0 to Y.0	Variance X.0, Variance Y.0	df_{num}, df_{den}	F_{obs}	$P(F \le F_{obs})$ or (1-p/2)	p-value
1.0 to 2.0	296.9, 256.5	18, 17	1.158	0.617	0.766
1.0 to 3.0	274.5, 315.9	22, 23	1.151	0.630	0.740
1.0 to 4.0	263.3, 261.9	24, 23	1.005	0.504	0.992
1.0 to 5.0	270.4, 248.5	22, 17	1.088	0.565	0.870
1.0 to 6.0	267.2, 66.5	22, 15	4.018	0.996	0.008*
1.0 to 7.0	194.6, 97.6	23, 22	1.994	0.945	0.110
2.0 to 3.0	258.1, 299.9	22, 21	1.162	0.633	0.734
2.0 to 4.0	253.2, 264.7	23, 21	1.045	0.538	0.924
2.0 to 5.0	226.8, 249.7	17, 20	1.101	0.586	0.828
2.0 to 6.0	264.2, 86.7	21, 16	3.047	0.987	0.026*
2.0 to 7.0	254.5, 100.6	22, 22	2.530	0.983	0.034*
3.0 to 4.0	317.4, 246.1	19, 19	1.290	0.708	0.584
3.0 to 5.0	256.7, 220.6	22, 17	1.164	0.620	0.760
3.0 to 6.0	307.9, 87.1	22, 17	3.535	0.995	0.010*
3.0 to 7.0	282.0, 73.2	20, 20	3.852	0.998	0.004*
4.0 to 5.0	263.6, 216.4	23, 17	1.218	0.657	0.686
4.0 to 6.0	272.1, 82.6	22, 16	3.294	0.991	0.018*
4.0 to 7.0	282.6, 101.9	22, 20	2.773	0.987	0.026*
5.0 to 6.0	223.8, 86.0	19, 17	2.602	0.974	0.052*
5.0 to 7.0	243.1, 105.7	16, 21	2.300	0.962	0.076*
6.0 to 7.0	88.6, 101.9	21, 16	1.150	0.607	0.786

* Significant at the 0.05 level (experiment-wise error rate). No significant results were seen at the 0.0024 level (comparison-wise error rate with the Bonferroni adjustment).

Table 7.

P-Values for Each H4 Pairwise Comparison

Category	2.0	3.0	4.0	5.0	6.0	7.0
1.0	0.766	0.740	0.992	0.870	0.008*	0.110
2.0		0.734	0.924	0.828	0.026*	0.034*
3.0			0.584	0.760	0.010*	0.004*
4.0				0.686	0.018*	0.026*
5.0					0.052	0.076
6.0						0.786

* Significant at the 0.05 level (experiment-wise error rate). No significant results were seen at the 0.0024 level (comparison-wise error rate with the Bonferroni adjustment).

Question 5 - How accurate are the evaluators' scores for each category?

Question 5 (Q5) examined the accuracy of the scores from a sample of untrained evaluators. Since this was the first known application of accuracy indices in this context, descriptive statistics were calculated for each index. Means and medians for each index by category are shown in Table 8. Q5 was addressed by graphical analysis and by testing hypothesis 5 (H5). Expert scores[3] for each item and category were plotted onto the box plots from H1 to facilitate comparison with the distribution of each group's scores (see Figure 6). Hypothesis 5 tested for differences between the mean accuracy of the control group and treatment group's scores. Since the data used for H5 were from the first evaluations (pre-treatment), no difference in accuracy was expected between the groups.

H5: There will be no difference in accuracy between the control group and treatment group during their first evaluations.

H5 was tested using accuracy statistics calculated from the control group's and treatment group's first evaluation scores. Elevation[4] and dimensional accuracy[5] (DA) were calculated for each evaluator for each category. Procedures for calculating these statistics were described in Appendix AT. Each group's elevation and DA indices for each category were plotted to check for normality. Since many of the plots appeared inconclusive, both t-tests and equivalent rank-based tests were performed for each category and index. The specific hypotheses tested to address H5 is shown below. A

[3] Expert scores were provided for each item as a ten point range. No information was provided regarding the exact number of experts involved or the actual distribution of the scores. For statistical purposes and graphical simplicity, the mean of this ten point range was used to estimate the true score.

[4] Elevation is the average of J's predictions over all items and ratees minus the central tendency of the self-descriptions (i.e., true scores) for all items and ratees combined (Cronbach, 1955, p. 178). Sulsky and Balzer's (1988) two-way ANOVA analogy described Elevation as the difference between the rater's grand mean and the true score grand mean.

[5] Dimensional accuracy (DA) is "equivalent to Cronbach's (1955) differential accuracy in that it measures the accuracy with which each rater evaluated a single ratee on each dimension" (Hauenstein and Alexander, 1991, p. 308). Dimensional accuracy is similar to McIntyre, Smith, and Hassett's correlational accuracy, in that it "measures the parallelism between subjects' scores and true scores" (1984, p. 151)

content analysis of the evaluators' comments[6] on strengths, areas for improvement, and site visit issues was used to compare each evaluator's comments to the experts' comments.

Table 8.

Group Mean and Median Accuracy Indices for each Category (1st Evaluation)

Category	Accuracy Index	Control Mean	Treatment Mean	Control Median	Treatment Median
1.0	Elevation	22.49	25.37	24.70	24.70
2.0	Elevation	22.78	19.25	28.35	15.00
3.0	Elevation	39.00	33.08	41.50	35.00
4.0	Elevation	18.80	18.09	18.80	18.80
5.0	Elevation	19.08	22.55	18.80	23.80
6.0	Elevation	28.67	26.64	31.70	30.00
7.0	Elevation	22.17	16.08	23.00	15.00
1.0	DA	9.85	10.26	8.50	10.30
2.0	DA	11.69	12.42	10.55	10.80
3.0	DA	7.82	8.08	5.00	5.00
4.0	DA	15.00	10.50	16.20	10.55
5.0	DA	18.11	14.13	17.50	16.10
6.0	DA	8.99	7.53	7.40	7.35
7.0	DA	11.03	8.40	9.70	7.60

Testing Hypothesis 5

T-tests were used to compare the accuracy of the control and treatment groups' scores on each category. For each category, separate tests were conducted using elevation and dimensional accuracy as indices of accuracy. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 1). A test-wise Type I error rate of alpha = 0.05 was used. Under the null hypotheses, it was

[6] These comments accompanied the score of each item and were intended to provide justification for the score.

reasonable to expect the variance of the accuracy indices to be the same; therefore, a pooled variance estimate was used.

Tested Hypotheses

H_0: There is no difference between the mean accuracy of Category X.0 scores from two randomly split-halves of a sample of untrained evaluators.
H_1: There is a difference between the mean accuracy of Category X.0 scores from two randomly split-halves of a sample of untrained evaluators.

Where X.0 = 1.0 to 7.0; and accuracy was measured using elevation for the first series of tests, then dimensional accuracy for the second series of tests. The corresponding hypotheses tested with the Mann-Whitney test actually tested the median accuracy rather than the mean accuracy.

Test Results

Distributions of both groups' dimensional accuracy and elevation for each category were plotted. The plots of some of the distributions did not appear to be normal. This is not surprising given the small sample sizes[7]. Both classical t-tests and Mann-Whitney tests (a.k.a. Wilcoxon rank sum tests) were performed for each category and accuracy index. The results are summarized in Table 9. The detailed output from the t-tests and Mann-Whitney tests are shown in Appendix AE.

[7] In some cases, a single data point made an otherwise normal looking distribution appear bi-modal.

Table 9.

Summary of Test Results for Hypothesis 5

Category	Accuracy Index	p-value (t-test)	p-value (Mann-Whitney)
1.0	Elevation	0.48	0.74
2.0	Elevation	0.50	0.44
3.0	Elevation	0.42	0.38
4.0	Elevation	0.84	0.96
5.0	Elevation	0.51	0.49
6.0	Elevation	0.65	0.67
7.0	Elevation	0.10	0.16
1.0	DA	0.83	0.74
2.0	DA	0.80	0.78
3.0	DA	0.90	0.76
4.0	DA	0.08	0.12
5.0	DA	0.27	0.29
6.0	DA	0.60	0.55
7.0	DA	0.15	0.16

Note: Group sample sizes for each category were listed in Table 2 and are listed again in Figure 6 (see next page).

Category 1.0 - Leadership

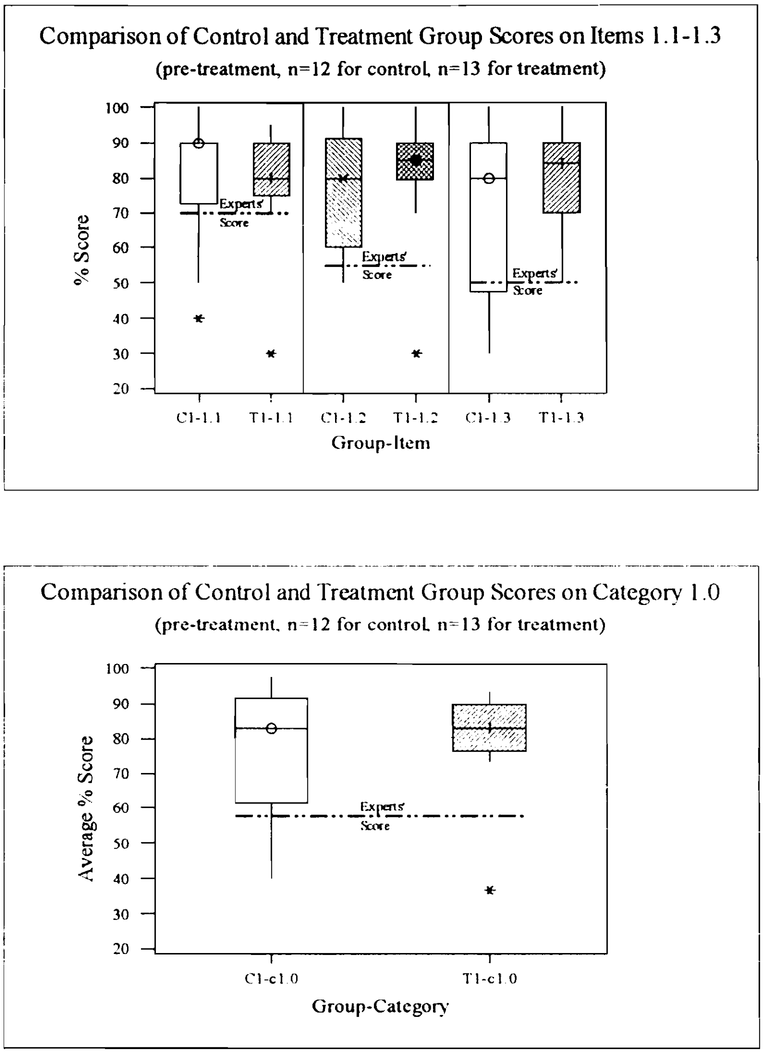


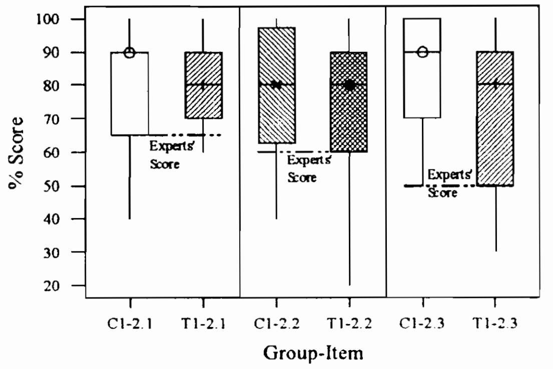
Figure 6. Comparison of expert scores to box plots of each group's scores (pre-treatment).

figure continues

Category 2.0 - Information and Analysis

Comparison of Control and Treatment Group Scores on Items 2.1-2.3

(pre-treatment, n=12 for control, n=11 for treatment)



Comparison of Control and Treatment Group Scores on Category 2.0

(pre-treatment, n=12 for control, n=11 for treatment)

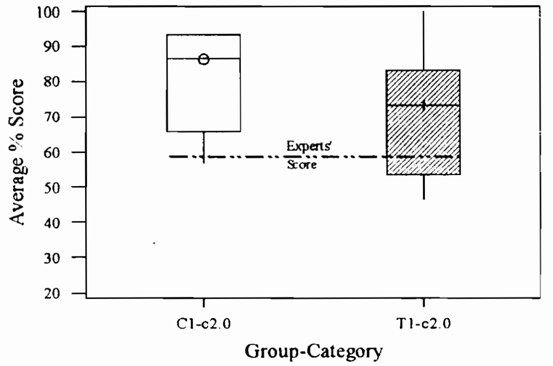
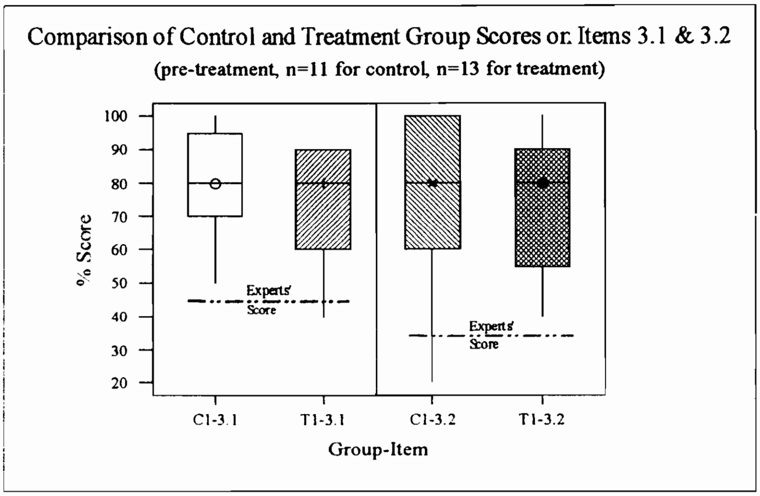


figure continues

Category 3.0 - Strategic Planning



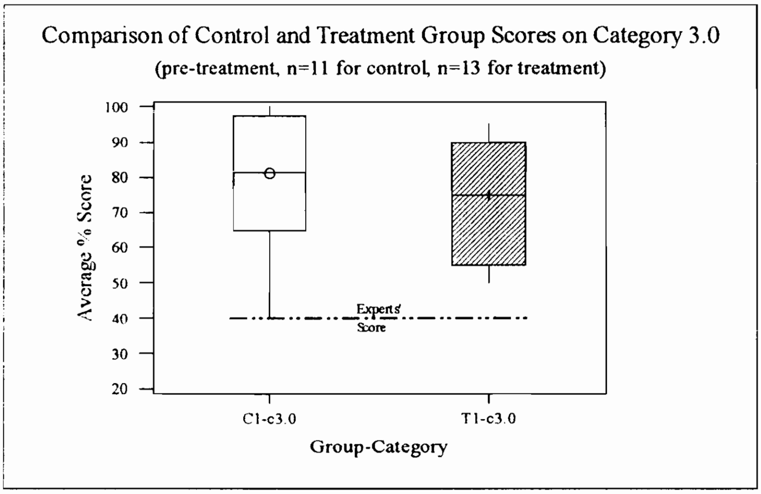
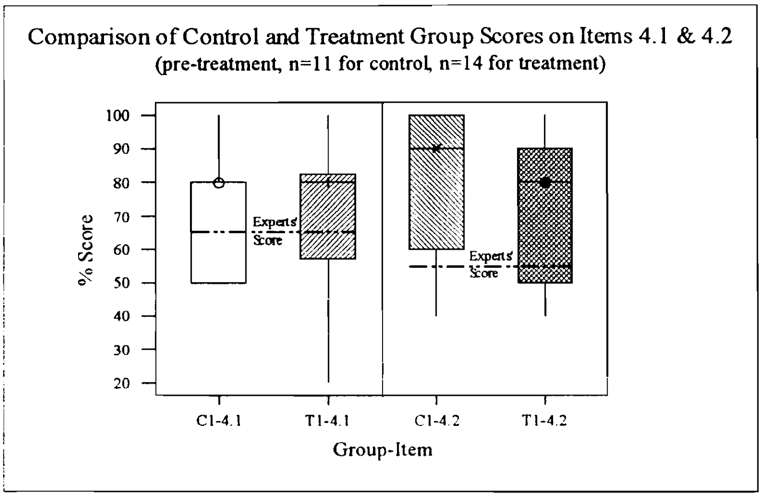


figure continues

Category 4.0 - Human Resource Development and Management


Comparison of Control and Treatment Group Scores on Items 4.1 & 4.2
(pre-treatment, n=11 for control, n=14 for treatment)

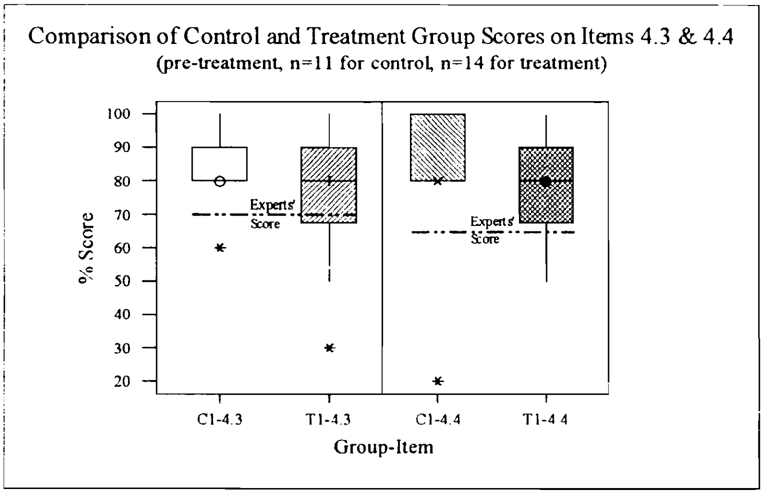
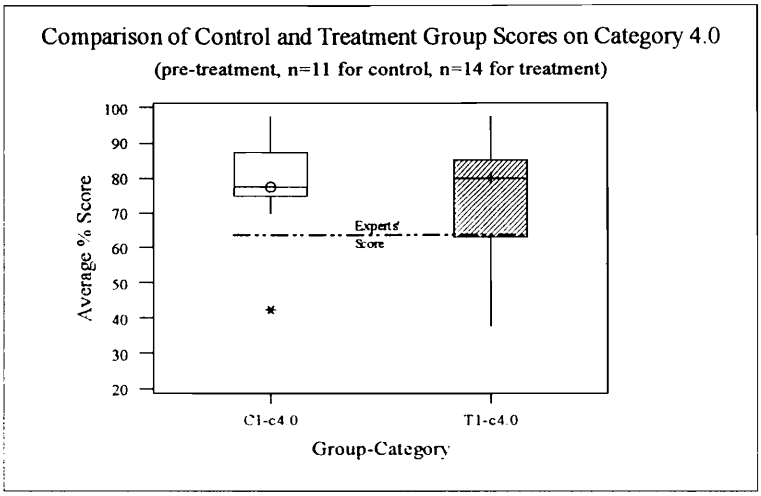
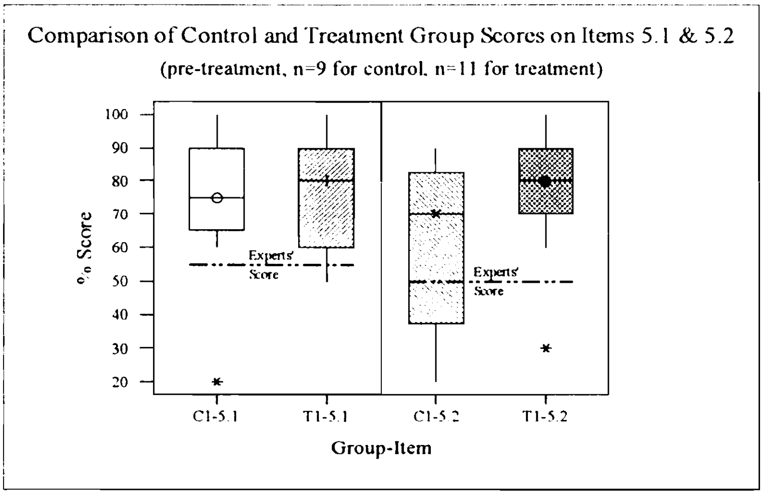

Comparison of Control and Treatment Group Scores on Items 4.3 & 4.4
(pre-treatment, n=11 for control, n=14 for treatment)

figure continues


Comparison of Control and Treatment Group Scores on Category 4.0

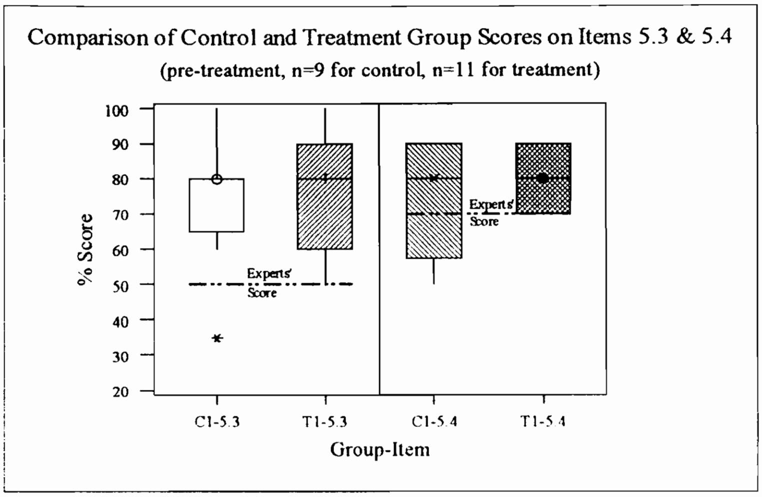
(pre-treatment, n=11 for control, n=14 for treatment)

Category 5.0 - Process Management


Comparison of Control and Treatment Group Scores on Items 5.1 & 5.2

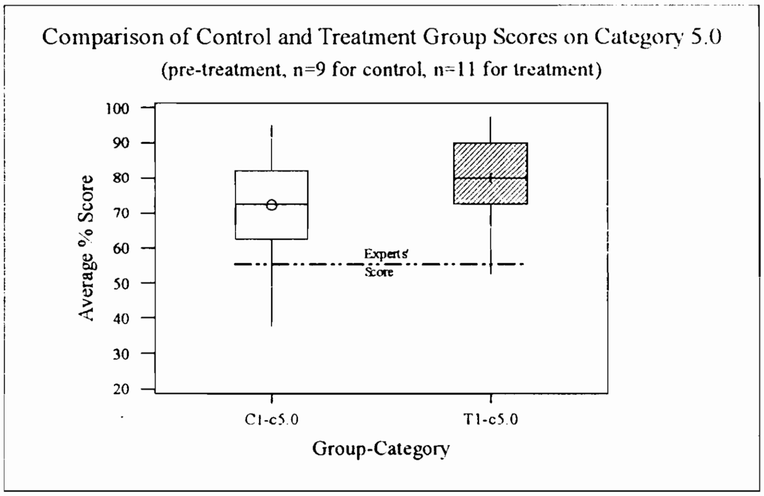
(pre-treatment, n=9 for control, n=11 for treatment)

figure continues



Comparison of Control and Treatment Group Scores on Items 5.3 & 5.4

(pre-treatment, n=9 for control, n=11 for treatment)

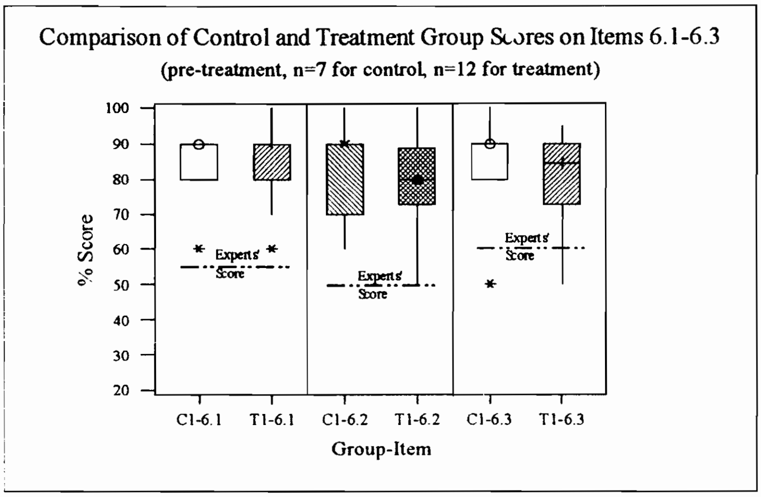


Comparison of Control and Treatment Group Scores on Category 5.0

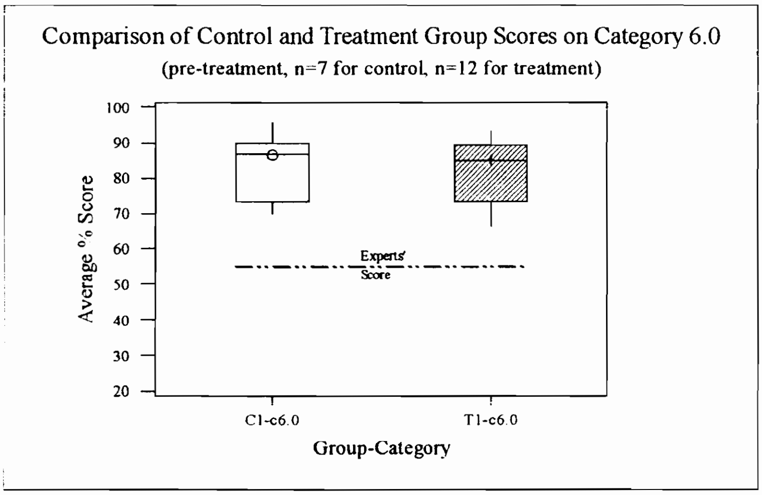
(pre-treatment, n=9 for control, n=11 for treatment)

figure continues

Category 6.0 - Business Results


Comparison of Control and Treatment Group Scores on Items 6.1-6.3

(pre-treatment, n=7 for control, n=12 for treatment)


Comparison of Control and Treatment Group Scores on Category 6.0

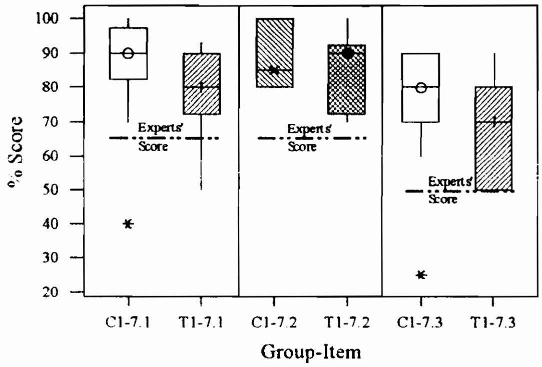
(pre-treatment, n=7 for control, n=12 for treatment)

figure continues

Category 7.0 - Customer Focus and Satisfaction

Comparison of Control and Treatment Group Scores on Items 7.1-7.3

(pre-treatment, n=12 for control, n=12 for treatment)



Comparison of Control and Treatment Group Scores on Items 7.4 & 7.5

(pre-treatment, n=12 for control, n=12 for treatment)

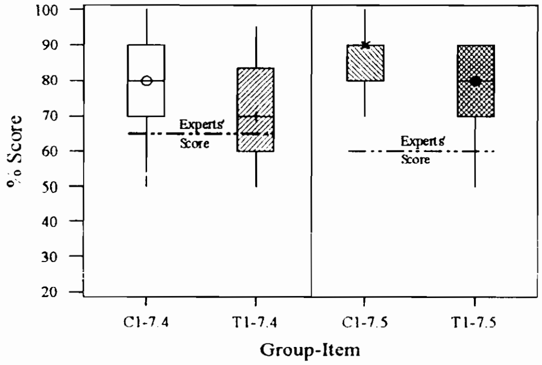
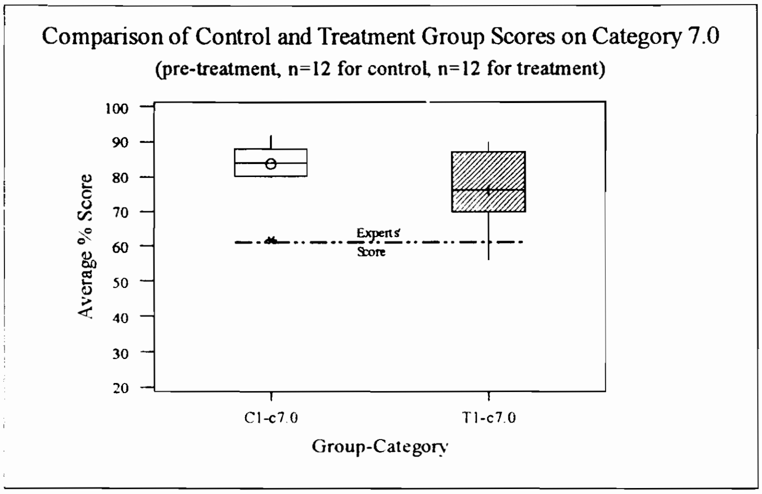


figure continues


Comparison of Control and Treatment Group Scores on Category 7.0
(pre-treatment, n=12 for control, n=12 for treatment)
Average % Score
100
90
80
70
60
50
40
30
20
Experts' Score
C1-c7.0
T1-c7.0
Group-Category

Content Analysis of First Evaluation Qualitative Comments

A content analysis of the evaluators' comments on strengths, areas for improvement, and site visit issues was used to compare the experts' comments to each evaluator's comments. These comments accompanied the score of each item and were intended to provide justification for the score. Each subject's comments were entered into a table to facilitate review and comparison. A separate table was constructed for each group. Each table was organized by item. That is, all the subjects' comments for a given item (e.g., 1.1) were listed consecutively, then the comments for the next (e.g., 1.2) item were listed. Each row in the table contains the comments of a single subject for that item. The first column is the subject number, the second column is the item number, followed by columns corresponding to the response sections of the comment and scoring worksheet[8]: (+/++), area to address, strengths; (-/--),area to address, areas for improvement; and site visit issues. Not all subjects provided input for each column; however, whatever was provided is listed in the table. An excerpt from the tables used to summarize these data is shown in Appendix AQ.

The content analysis followed the procedure described below.

1. Prior to analyzing an item, the experts' comments for that item were read and the text of the case study reviewed. These had been carefully read and highlighted during the grading process.
2. The strengths identified by the first subject were read.
3. The experts' list of strengths was reviewed one-by-one. As each experts' strength was reviewed, it was compared to the subject's strengths to see if any of the subject's comments appeared to match it in wording or intent. If one of the subject's strengths appeared to match the experts' strength, it was labeled a "hit." Once an experts' strength had been matched or hit, any other matching strengths by this subject were labeled redundant[9]. Then the next strength from the experts' list

[8] See Appendix K for an example of the comment and scoring worksheet. Subjects were provided scorebooks containing one of these worksheets for each item. On the worksheets, (+/++) and area to address were spaces provided to note the relative importance and specific area of the item (e.g., a, b, c) to which each strength applies. (-/--) and area to address were similar spaces provided for each area for improvement.

[9] Differing levels of aggregation were handled by deeming the experts' comments as the master. That is, multiple comments by a subject may pertain to one of the experts' comments, but only be counted as one

was reviewed and compared to the subjects' strengths. To accommodate the subjectivity inherent in this process, comments that appeared to match but left some doubt in the analyst's opinion were labeled as "almost hits."

4. Comments by the subject that appeared unrelated to the experts' list of strengths were labeled as "misses." As before, if multiple comments appeared to address the same issue, all but one of these were labeled as redundant.
5. After reviewing all the strengths in the experts' list, the number of hits (H), almost hits (A), and misses (M) for that subject were counted. The number of redundant comments was not counted.
6. Steps 2 through 5 were repeated for the areas for improvement.
7. Steps 2 through 5 were repeated for site visit issues.
8. The counts from steps 5, 6, and 7 were entered into a spreadsheet for later analysis.
9. Steps 2 through 8 were repeated for the next subject until all the subjects in that group who had evaluated this item were analyzed.
10. Steps 1 through 9 were repeated for the next item.

Appendix AQ contains examples of two subjects' (#1446 and #1513) comments for Item 1.1 in the tabular format described earlier. The experts' comments for the same Item are contained in Appendix AR. Subject #1446's list of strengths matched three of the nine strengths the experts identified, thus subject #1446 was credited with three hits (H = 3) Subject #1446 identified a fourth strength that may match one of the experts' strengths and was credited with one almost hit (A = 1). Subject #1446 offered no other comments regarding strengths and was credited with zero misses (m = 0). With respect to strengths, Subject #1446 had three hits, one almost hit, and no misses compared to the experts' list of nine strengths. With respect to areas for improvement, Subject #1446 had one hit, one almost hit, and two misses compared to the experts' list of four areas for improvement. Subject #1446 suggested no site visit issues, this was counted as a null response. When a null response was entered into the summary spreadsheet, it did not effect or skew the

hit. If a single comment by a subject applied to more than one of the experts' comments, if could be labeled as more than one hit.

calculation of averages (e.g., average number of hits). Null responses[10] were reflected in the summary statistic %Experts, as explained below. For comparison purposes, Subject #1513 had three hits, no almost hits, and 1 miss with respect to strengths on Item 1.1. Subject #1513 had zero hits, 1 almost hit, and 1 miss with respect to areas for improvement and no hits, one almost hit, and no misses with respect to site visit issues.

The count data for hits, almost hits, and misses are contained in Appendix AS. Means were calculated for the hits, almost hits, and misses for strengths, areas for improvement, and site visit issues for each item. These means were calculated for each group and for the combined sample (i.e., the baseline for future comparison). Summary statistics were calculated for strengths and for all comments (i.e., strengths, areas for improvement, and site visit issues). These summary statistics were called %Hits (%H) and %Experts (%Exp).

%Hits was the number of hits divided by the number of counted comments[11] a subject submitted. For example, subject #1446 submitted four strengths for Item 1.1 and three of these were hits. This gave subject #1446 a %Hits of 3/4 or 75% for Item 1.1 strengths. %Hits was not calculated for areas for improvement and site visit issues, because of the low average number of hits for these comments. Instead, %Hits was calculated for the combination of strengths, areas for improvement, and site visit issues. For example subject #1446 had three hits on strengths, one hit on areas for improvement, no hits on site visit issues out of eight total comments. This gave subject #1446 an overall %Hits of 4/8 or 50% for Item 1.1. %Hits indicates what proportion of a subject's comments were on target. A subject offering few comments, but with most of them matching the experts' comments would result in a high %Hits. Another subject matching the same number of the experts' comments, but also offering a lot of other comments (e.g., misses) would result in a lower %Hits. %Hits might be viewed as a measure of the

[10] Nearly every subject with usable scores submitted at least one comment for strengths on each item. Null responses were occasionally submitted for areas for improvement and frequently submitted for site visit issues.

[11] Redundant comments were not counted and therefore. not included in the denominator for %Hits.

quality of the subject's comments. Unfortunately, a subject may offer only one comment which was an obvious strength or area for improvement and receive a %Hits of 100%. The second summary statistic, %Experts, addresses this weakness.

%Experts was the proportion of the experts' comments that the subject matched. %Experts was calculated by adding the number of hits and 50% times the number of almost hits, then dividing this sum by the number of comments the experts offered. Like %Hits, %Experts was calculated for both strengths and the combination of strengths, areas for improvement, and site visit issues. Null responses were reflected in the calculation of %Experts since the denominator counts all the experts' comments. A subject may not have submitted any comments for site visit issues, but their total hits and almost hits were still divided by a sum that included the number of site visit issues the experts identified.

The calculation of %Experts is illustrated in the following example. Subject #1446 had three hits and one almost hit on Item 1.1 strengths. Counting the one almost hit as 50% of a hit, this results in a %Experts of 3.5/9 or 39% of Item 1.1 strengths. Similarly, Subject #1446 had four hits and two almost hits for all comments on Item 1.1, compared to 18 comments[12] by the experts. This gave subject #1446 a %Experts of 5/18 or 28% on Item 1.1. %Experts might be viewed as a measure of the effectiveness of a subject's comments. The hypothetical subject with only one comment (a hit) would likely receive a low %Experts. A subject with the same number of hits and almost hits, but with many more comments would receive the same %Experts and a lower %Hits.

Table 10 summarizes the mean number of hits, almost hits, misses, %Hits, and %Experts for the combined control and treatment groups. The mean counts and statistics for the control and treatment groups are given in Appendix AS.

[12] The experts identified nine strengths, four areas for improvement, and five site visit issues for Item 1.1.

Table 10.

Summary of Count Data and Statistics from the Q5 Content Analysis (1st Evaluation)

		MEANS				
Item	Type of Comment	Hits	Almost Hits	Misses	%Hits	%Experts
1.1						
	Strengths	3.6	1.5	1.2	58%	48% (of 9)[13]
	Areas for Improvement	0.6	0.5	1.6		(4)
	Site Visit Issues	0.6	0.5	1.0		(5)
	Overall				47%	30% (of 18)
1.2						
	Strengths	2.5	0.7	0.3	71%	56% (of 5)
	Areas for Improvement	0.6	0.6	0.7		(6)
	Site Visit Issues	0.8	0.5	0.3		(5)
	Overall				57%	26% (of 16)
1.3						
	Strengths	2.9	0.2	1.0	73%	59% (of 5)
	Areas for Improvement	0.8	0.3	1.2		(6)
	Site Visit Issues	0.3	0.4	1.0		(5)
	Overall				58%	25% (of 16)
2.1						
	Strengths	2.0	0.8	3.3	34%	48% (of 5)
	Areas for Improvement	0.5	0.2	1.5		(4)
	Site Visit Issues	0.2	0.3	1.0		(3)
	Overall				30%	25% (of 12)
2.2						
	Strengths	1.2	0.5	2.5	30%	48% (of 3)
	Areas for Improvement	0.4	0.1	2.2		(3)
	Site Visit Issues	0.3	0.6	1.0		(3)
	Overall				26%	22% (of 9)
2.3						
	Strengths	1.1	0.8	1.8	26%	38% (of 4)
	Areas for Improvement	0.7	0.3	0.9		(3)
	Site Visit Issues	0.1	0.1	0.9		(4)
	Overall				31%	20% (of 11)

table continues

[13] The numbers in parentheses represent the number of strengths, areas for improvement, or site visit issues the experts identified.

		MEANS				
Item	Type of Comment	Hits	Almost Hits	Misses	%Hits	%Experts
3.1						
	Strengths	1.3	0.5	1.7	36%	38% (of 4)
	Areas for Improvement	0.5	0.4	1.6		(6)
	Site Visit Issues	0.2	0.1	1.1		(6)
	Overall				30%	15% (of 16)
3.2						
	Strengths	0.7	0.8	1.2	28%	28% (of 4)
	Areas for Improvement	1.0	0.6	1.1		(4)
	Site Visit Issues	0.4	0.2	0.9		(4)
	Overall				31%	19% (of 12)
4.1						
	Strengths	0.5	0.4	4.7	10%	23% (of 3)
	Areas for Improvement	0.5	0.2	2.3		(5)
	Site Visit Issues	0.1	0.3	1.4		(5)
	Overall				11%	10% (of 13)
4.2						
	Strengths	0.9	0.6	1.7	31%	30% (of 4)
	Areas for Improvement	0.2	0.2	1.6		(3)
	Site Visit Issues	0.2	0.4	0.8		(4)
	Overall				23%	15% (of 11)
4.3						
	Strengths	0.9	1.0	2.2	25%	28% (of 5)
	Areas for Improvement	0.2	0.4	1.3		(3)
	Site Visit Issues	0.1	0.5	0.8		(6)
	Overall				19%	14% (of 14)
4.4						
	Strengths	1.6	0.3	1.8	42%	44% (of 4)
	Areas for Improvement	0.4	0.3	1.1		(4)
	Site Visit Issues	0.4	0.1	1.0		(4)
	Overall				35%	20% (of 12)
5.1						
	Strengths	1.0	0.8	3.7	18%	36% (of 4)
	Areas for Improvement	0.2	0.3	2.4		(4)
	Site Visit Issues	0.3	0.0	1.2		(5)
	Overall				15%	15% (of 13)

table continues

		MEANS				
Item	Type of Comment	Hits	Almost Hits	Misses	%Hits	%Experts
5.2						
	Strengths	0.3	0.6	3.5	7%	15% (of 4)
	Areas for Improvement	0.1	0.5	1.9		(3)
	Site Visit Issues	0.1	0.0	1.3		(5)
	Overall				6%	9% (of 12)
5.3						
	Strengths	0.6	0.7	4.4	10%	25% (of 4)
	Areas for Improvement	0.5	0.7	1.7		(4)
	Site Visit Issues	0.1	0.1	1.0		(4)
	Overall				12%	15% (of 12)
5.4						
	Strengths	1.8	0.5	2.7	43%	40% (of 5)
	Areas for Improvement	0.0	0.0	1.9		(1)
	Site Visit Issues	0.3	0.3	1.0		(6)
	Overall				28%	18% (of 12)
6.1						
	Strengths	1.5	0.4	0.1	80%	41% (of 4)
	Areas for Improvement	0.3	0.1	1.4		(4)
	Site Visit Issues	0.0	0.0	1.4		(4)
	Overall				49%	16% (of 12)
6.2						
	Strengths	1.1	0.7	0.1	54%	45% (of 3)
	Areas for Improvement	0.3	0.1	2.4		(4)
	Site Visit Issues	0.1	0.4	1.1		(4)
	Overall				28%	17% (of 11)
6.3						
	Strengths	1.0	0.8	1.1	33%	28% (of 5)
	Areas for Improvement	0.2	0.2	1.5		(3)
	Site Visit Issues	0.0	0.3	0.7		(4)
	Overall				23%	14% (of 12)
7.1						
	Strengths	1.2	1.1	2.7	27%	35% (of 5)
	Areas for Improvement	0.3	0.2	1.9		(3)
	Site Visit Issues	0.1	0.0	1.4		(4)
	Overall				22%	18% (of 12)

table continues

Item	Type of Comment	MEANS				
		Hits	Almost Hits	Misses	%Hits	%Experts
7.2						
	Strengths	1.4	1.2	3.0	23%	40% (of 5)
	Areas for Improvement	0.1	0.1	2.0		(3)
	Site Visit Issues	0.0	0.0	1.6		(5)
	Overall				20%	16% (of 13)
7.3						
	Strengths	1.1	0.9	1.4	33%	32% (of 5)
	Areas for Improvement	0.3	0.2	1.6		(6)
	Site Visit Issues	0.2	0.2	1.2		(4)
	Overall				25%	13% (of 15)
7.4						
	Strengths	0.9	0.7	1.1	32%	25% (of 5)
	Areas for Improvement	0.2	0.1	1.3		(4)
	Site Visit Issues	0.0	0.3	1.1		(2)
	Overall				23%	13% (of 11)
7.5						
	Strengths	0.4	0.1	1.5	22%	14% (of 3)
	Areas for Improvement	0.1	0.4	1.8		(5)
	Site Visit Issues	0.0	1.0	0.5		(4)
	Overall				18%	6% (of 12)

For a given sample of untrained evaluators, will evaluator training change the consistency of their scores?

This broad question was addressed by comparing the pre-training versus post-training scores for the control and treatment groups and by testing hypothesis 6 (H6). The comparison of pre-training versus post-training scores was done by calculating descriptive statistics and constructing box plots for each item and category. Addressing research questions 6 through 9 also contributed to this comparison. The descriptive statistics are shown in Appendix AF. The box plots are shown in Figure 7. Each box plot shows the distribution of pre- and post-training scores of the control group beside the distribution of pre- and post-training scores of the treatment group.

H6: There will be a difference in scores between the control and treatment groups during the second evaluation: a) by item, and b) by category.

H6 is the post treatment (post-training) counterpart to H1, which compared mean scores of the control and treatment groups prior to the training. Unlike H1, H6 is expected to show a difference between the control and treatment groups due to the effect of the training. The specific hypotheses tested to address H6 are described below.

Testing Hypothesis 6

A two-way (group x time) analysis of variance (ANOVA) was used to test for an interaction between group (untrained versus trained) and time (first evaluation versus second evaluation[14]). If no interaction effect was seen, group (row) and time (column) effects were examined. If mild ($p \leq 0.10$) or stronger evidence[15] of an interaction was present, simple group and time effects were tested. When there was little evidence of an interaction effect, but mild ($p \leq 0.10$) or stronger evidence of a main effect, the simple effects related to that main effect were tested. T-tests were used for all simple effects

[14] First evaluation and second evaluation might also be viewed as pre-treatment and post-treatment.

[15] This lower standard of evidence ($p \leq 0.10$) is only used to determine if further analysis is warranted, not to determine statistical significance.

tests. Although Minitab was used to calculate actual p-values, a test-wise Type I error rate of alpha = 0.05 was used to identify significant results. The F-statistics for the two-way ANOVA were calculated using the general linear model (GLM). The GLM was used due to the unbalanced design. Under the null hypotheses of the t-tests, it was reasonable to expect the variance of the scores to be the same; therefore, a pooled variance estimate was used. Prior to analysis, the data were screened as described below.

Data Screening

Those subjects participating in both courses were identified and all their responses from the second course were discarded. Their evaluations from the second course were likely affected by their experience in the earlier course. Their second evaluations from the first course were also discarded. The earliest submission date for a second evaluation in the first course was two days after distribution of the first evaluation materials in the second course. This overlap may have provided additional experience to those in both courses; therefore, their second evaluations in the first course may have been contaminated. Since paired comparisons were not being performed, these second evaluations were independent and should have no impact on the validity of the scores from their first evaluations. If a subject's scores for the items of a category were incomplete, then all the scores for that category were discarded. If a subject did not complete the first evaluation, their second evaluation scores were discarded due to potential loss of learning effect. Any subject's scores suspected of contamination due to missing submittal deadlines or improper exposure to the treatment were also discarded. The screened data (item scores) for the control and treatment groups' second evaluation are shown in Appendices BC and BD, respectively.

Tested Hypotheses

The group x time ANOVA tested the following hypotheses for each item and each category.

H_0: There are no interaction effects when comparing evaluators' mean scores based on both group (urtrained versus trained) and time (first evaluation versus second evaluation).
H_0: There are no group effects when comparing evaluators' mean scores averaged over time (first evaluation and second evaluation).
H_0: There are no time effects when comparing evaluators' mean scores averaged across groups (untrained and trained).
H_1: There is an effect (interaction, group, or time).

The following hypotheses were tested using a t-test when simple group effects were examined. These tests were run using second evaluation (post-treatment) data, since the same hypothesis was tested for H1 using first evaluation (pre-treatment) data.

H_0: There is no difference between the Item X.X mean scores of a trained group of evaluators and an untrained group of evaluators.
H_1: There is a difference between the Item X.X mean scores of a trained group of evaluators and an untrained group of evaluators.
where X.X = 1.1 to 7.5 for items or X.X = 1.0 to 7.0 for categories.

The following hypotheses were tested using a t-test when simple time effects were examined. These tests were run using data from both the control and treatment groups.

H_0: There is no difference between the first evaluation X.X mean scores and the second evaluation X.X mean scores for untrained evaluators.
H_1: There is a difference between the first evaluation X.X mean scores and the second evaluation X.X mean scores for untrained evaluators.
where X.X = 1.1 to 7.5 for items or X.X = 1.0 to 7.0 for categories.

H_0: There is no difference between the first evaluation X.X mean scores and the second evaluation X.X mean scores for trained evaluators.
H_1: There is a difference between the first evaluation X.X mean scores and the second evaluation X.X mean scores for trained evaluators.
where X.X = 1.1 to 7.5 for items or X.X = 1.0 to 7.0 for categories.

Test Results

The results of the two-way ANOVAs for Items 1.1 through 7.5 are summarized in Table 11. The results of the two-way ANOVAs for Categories 1.0 through 7.0 are summarized in Table 12. Complete ANOVA tables for each item and category are shown in Appendix AG. The results of the simple effects t-tests for items and categories are also summarized in Tables 11 and 12, respectively, along with the evidence of an interaction or main effect that led to the t-test.

Table 11. Summary of ANOVA and t-test Results for Hypothesis 6 by Item.

Item	Interaction Effect	Main Effects		Simple Effects			
		Group	Time	Group		Time	
				Pre-	Post-	Control	Treatment
1.1	0.581	0.411	0.978				
1.2	0.393	0.782	0.840				
1.3	0.144	0.892	0.785				
2.1	0.656	0.462	0.064*			0.41	0.042**
2.2	0.530	0.071*	0.163	0.38	0.078		
2.3	0.449	0.291	0.046**			0.066	0.36
3.1	0.780	0.210	0.208				
3.2	0.632	0.202	0.100*			0.43	0.12
4.1	0.246	0.244	0.084*			0.018**	0.71
4.2	0.396	0.176	0.744				
4.3	0.953	0.071*	0.215	0.24	0.13		
4.4	0.854	0.287	0.486				
5.1	0.314	0.672	0.263				
5.2	0.057*			0.20	0.16	0.063	0.61
5.3	0.053*			0.51	0.035**	0.18	0.18
5.4	0.013**			0.27	0.02**	0.063	0.12
6.1	0.087*			0.95	0.043**	0.15	0.0006**
6.2	0.032**			0.69	0.0003**	0.40	0.0009**
6.3	0.103	0.044**	0.054*	0.77	0.016**	0.81	0.022**
7.1	0.933	0.248	0.405				
7.2	0.705	0.466	0.616				
7.3	0.609	0.042**	0.244	0.30	0.03**		
7.4	0.467	0.024**	0.422	0.14	0.11		
7.5	0.564	0.012**	0.244	0.085	0.085		

* = mild evidence of an interaction or main effect, $p \le 0.10$
** = evidence of significant effect, $p \le 0.05$

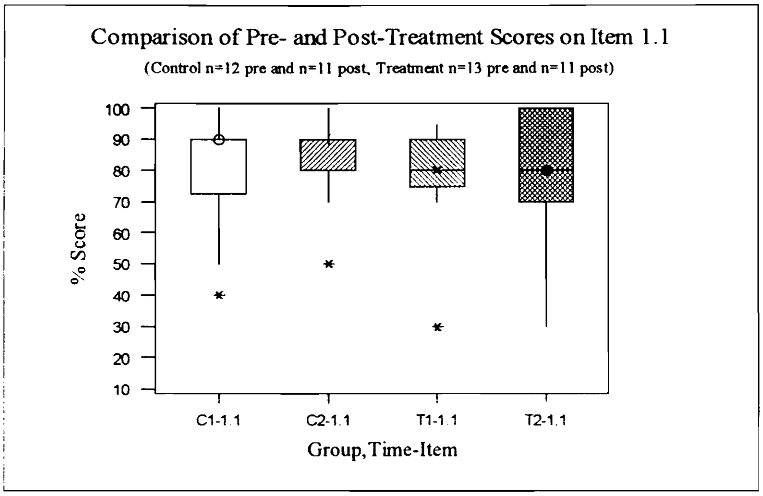
Table 12. Summary of ANOVA and t-test Results for Hypothesis 6 by Category.

Item or	Interaction	Main Effects		Simple Effects			
Category	Effect	Group	Time	Group		Time	
				Pre-	Post-	Control	Treatment
1.0	0.274	0.706	0.850				
2.0	0.906	0.134	0.035**			0.19	0.094
3.0	0.670	0.165	0.107				
4.0	0.470	0.113	0.245				
5.0	0.036**			0.31	0.051	0.066	0.37
6.0	0.028**			0.65	0.0038*	0.29	0.0005**
7.0	0.664	0.015**	0.558	0.084	0.098		

* = mild evidence of an interaction or main effect, $p \leq 0.10$

** = evidence of significant effect, $p \leq 0.05$

Category 1.0 - Leadership



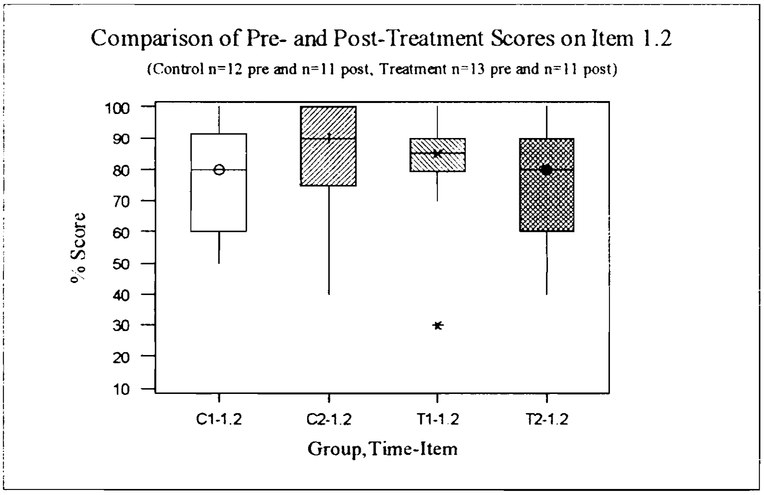
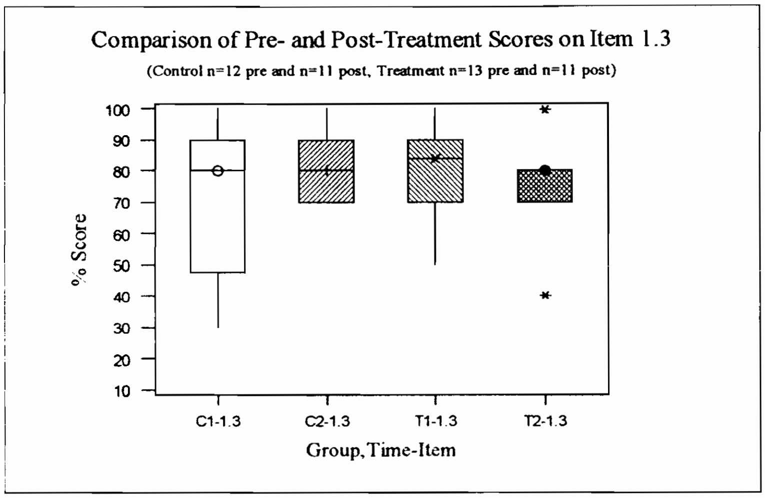


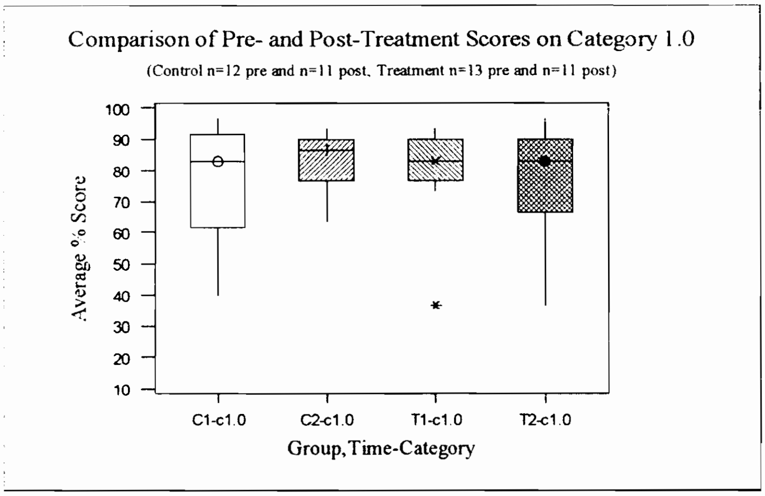
Figure 7. Box plot comparisons of pre- and post-treatment scores.

figure continues



Comparison of Pre- and Post-Treatment Scores on Item 1.3

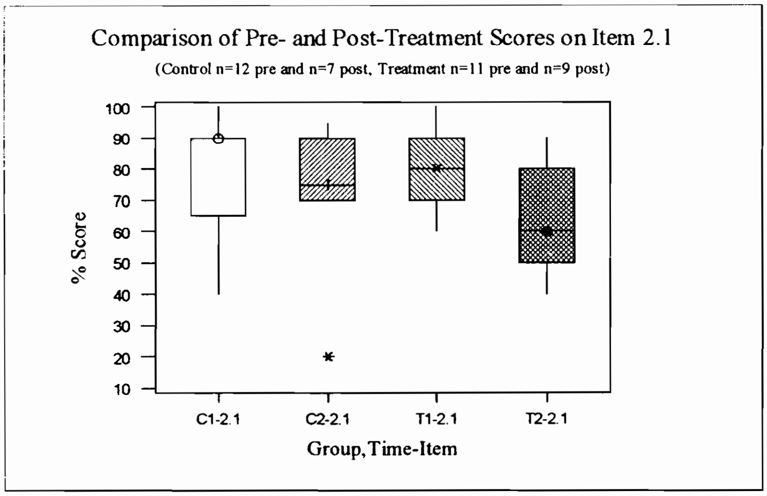
(Control n=12 pre and n=11 post, Treatment n=13 pre and n=11 post)



Comparison of Pre- and Post-Treatment Scores on Category 1.0

(Control n=12 pre and n=11 post, Treatment n=13 pre and n=11 post)

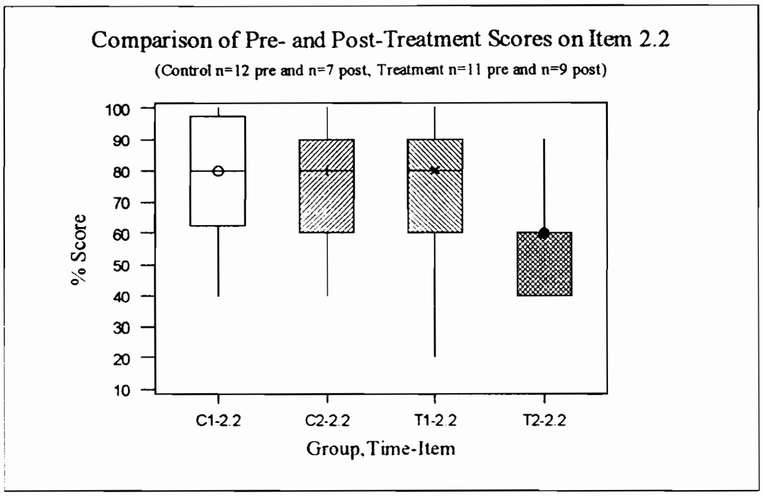
Category 2.0 - Information and Analysis



Comparison of Pre- and Post-Treatment Scores on Item 2.1

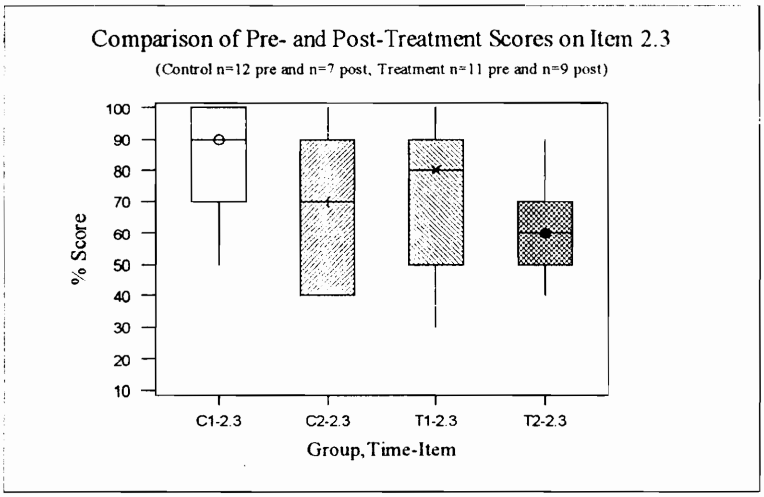
(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)

figure continues



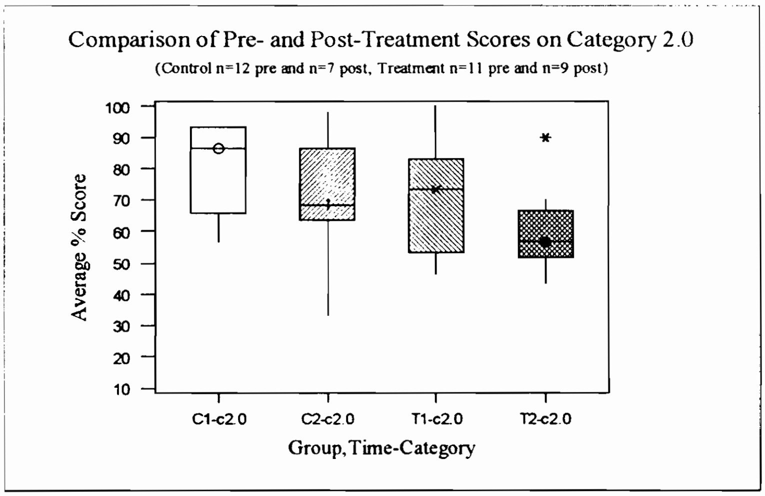
Comparison of Pre- and Post-Treatment Scores on Item 2.2

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 2.3

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 2.0

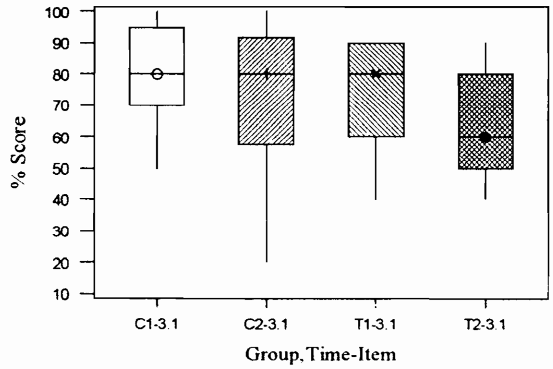
(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)

figure continues

Category 3.0 - Strategic Planning

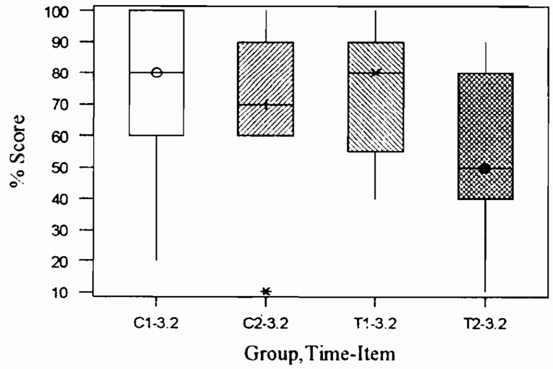
Comparison of Pre- and Post-Treatment Scores on Item 3.1

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 3.2

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 3.0

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)

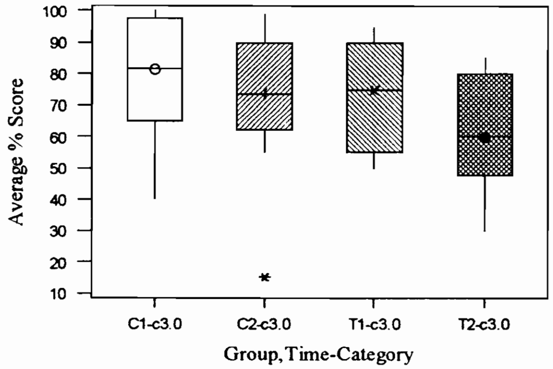
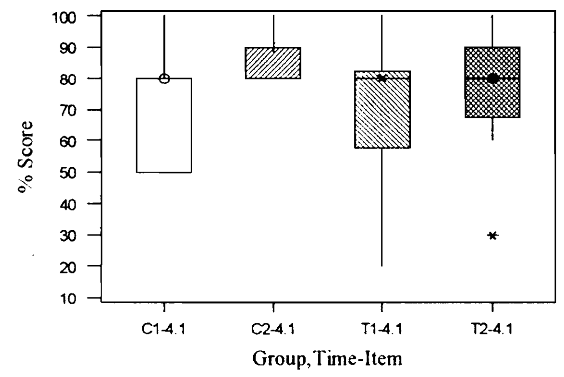
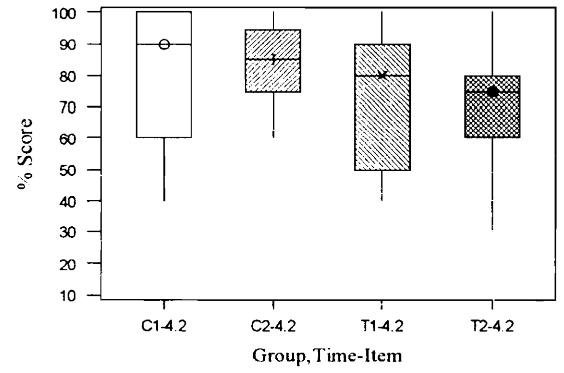


figure continues

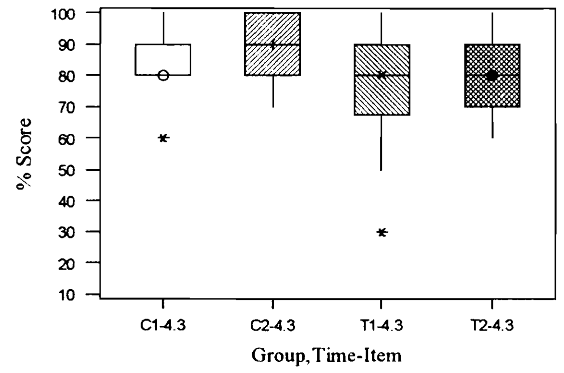
Category 4.0 - Human Resource Development and Management


Comparison of Pre- and Post-Treatment Scores on Item 4.1

(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)

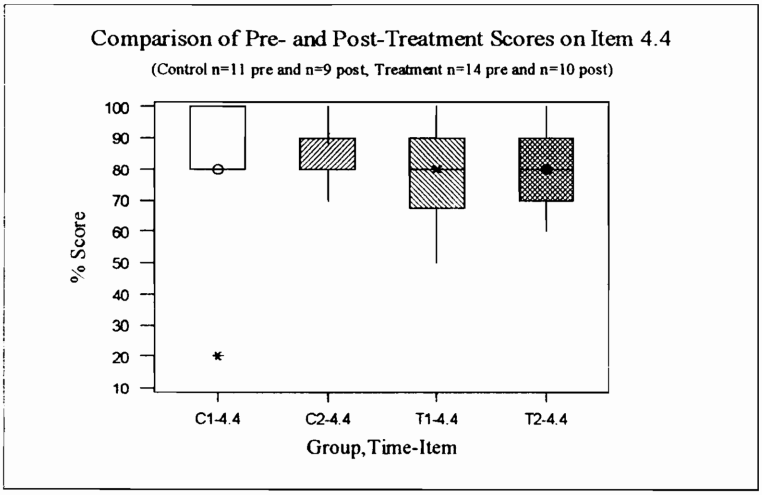

Comparison of Pre- and Post-Treatment Scores on Item 4.2

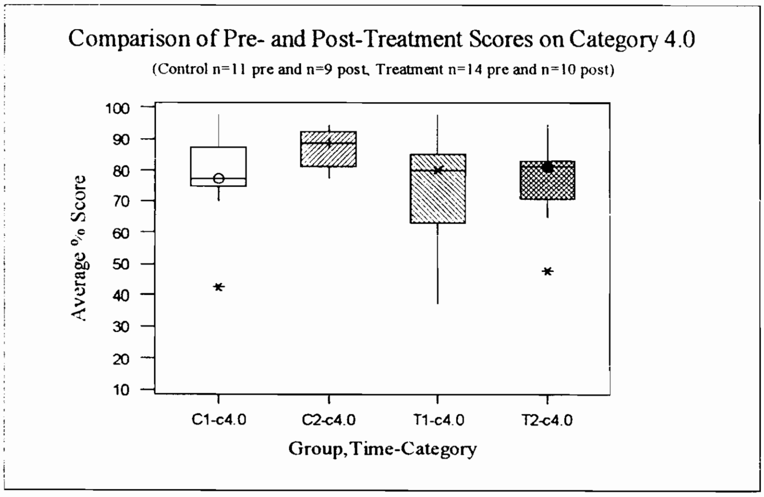
(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)


Comparison of Pre- and Post-Treatment Scores on Item 4.3

(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)

figure continues





Category 5.0 - Process Management

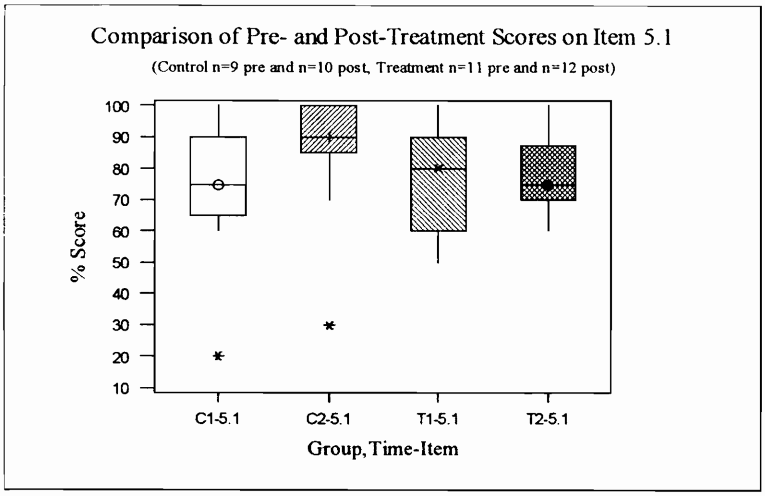
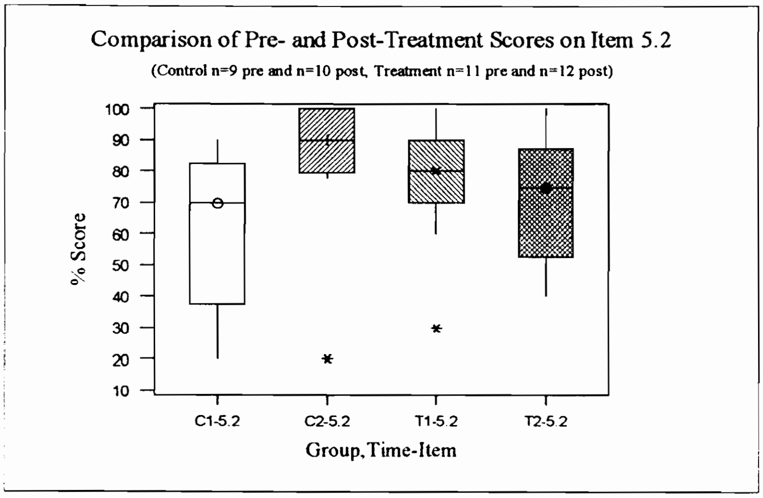
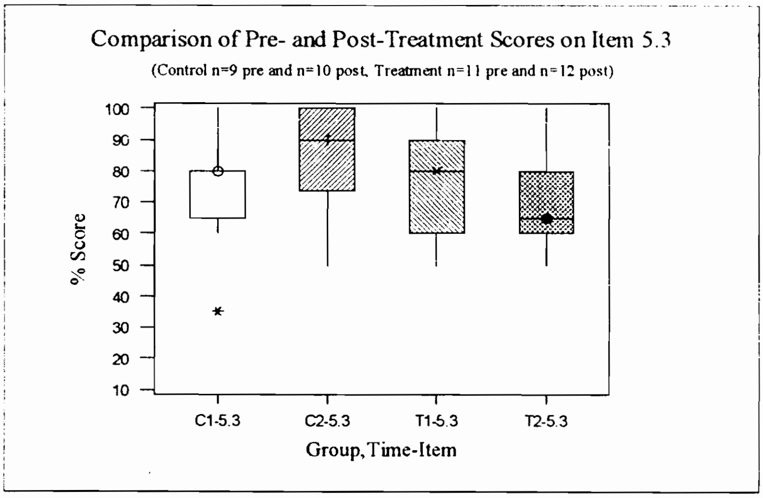


figure continues



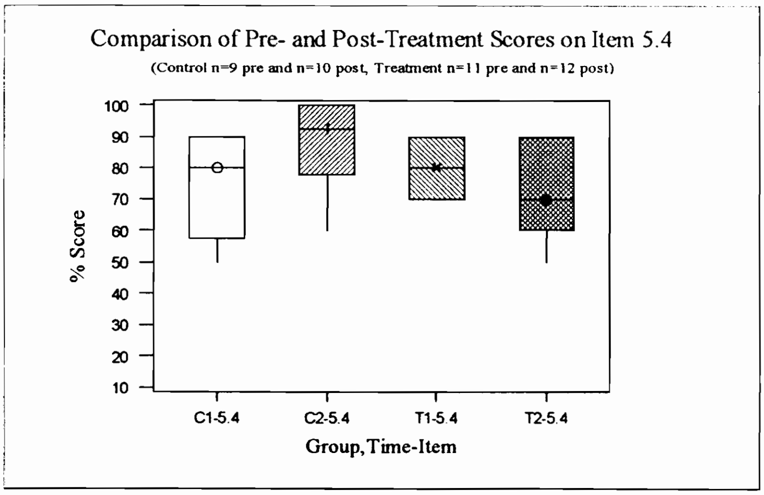
Comparison of Pre- and Post-Treatment Scores on Item 5.2

(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)



Comparison of Pre- and Post-Treatment Scores on Item 5.3

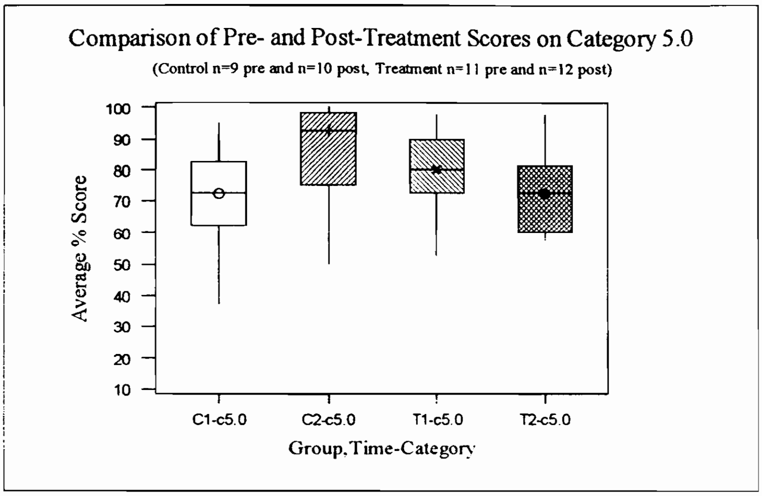
(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)



Comparison of Pre- and Post-Treatment Scores on Item 5.4

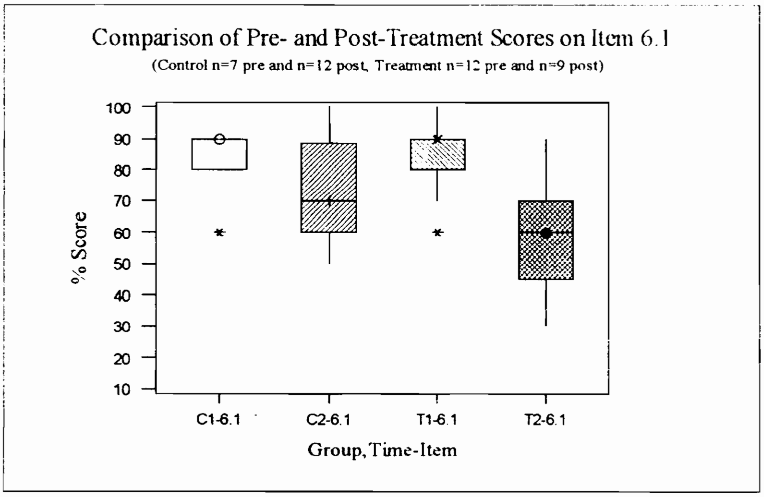
(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)

figure continues

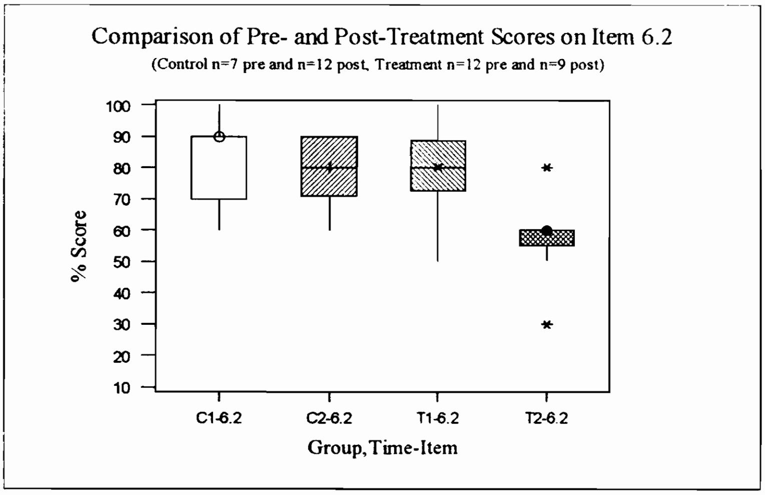

Comparison of Pre- and Post-Treatment Scores on Category 5.0

(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)

Category 6.0 - Business Results

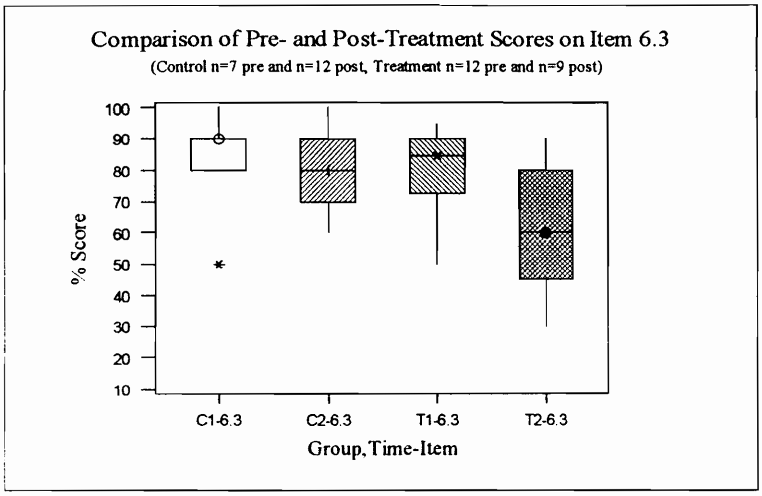

Comparison of Pre- and Post-Treatment Scores on Item 6.1

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)


Comparison of Pre- and Post-Treatment Scores on Item 6.2

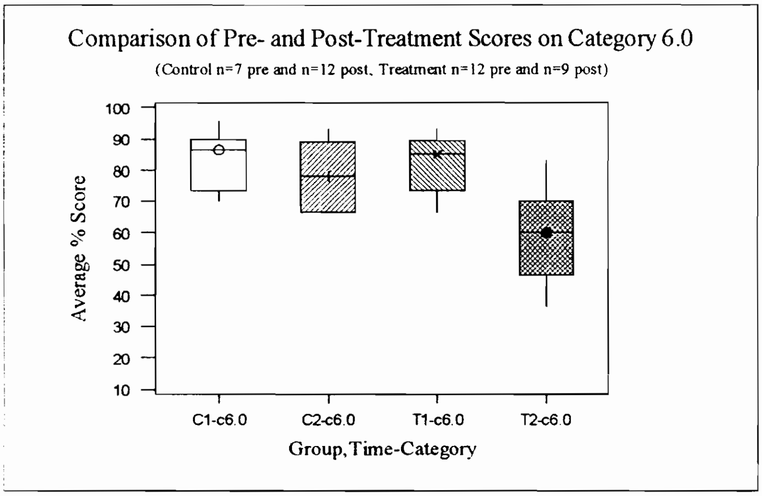
(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)

figure continues



Comparison of Pre- and Post-Treatment Scores on Item 6.3

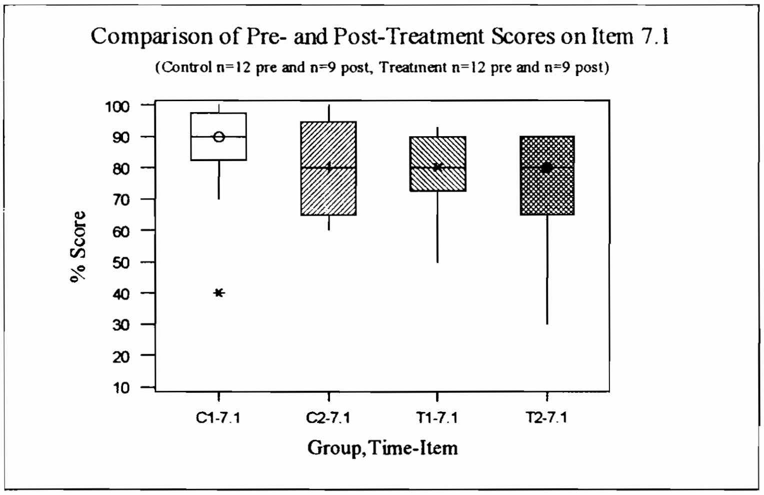
(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 6.0

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)

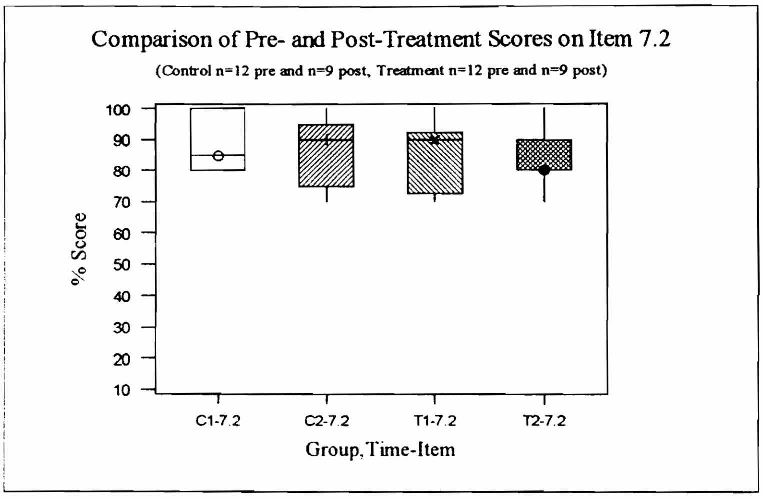
Category 7.0 - Customer Focus and Satisfaction



Comparison of Pre- and Post-Treatment Scores on Item 7.1

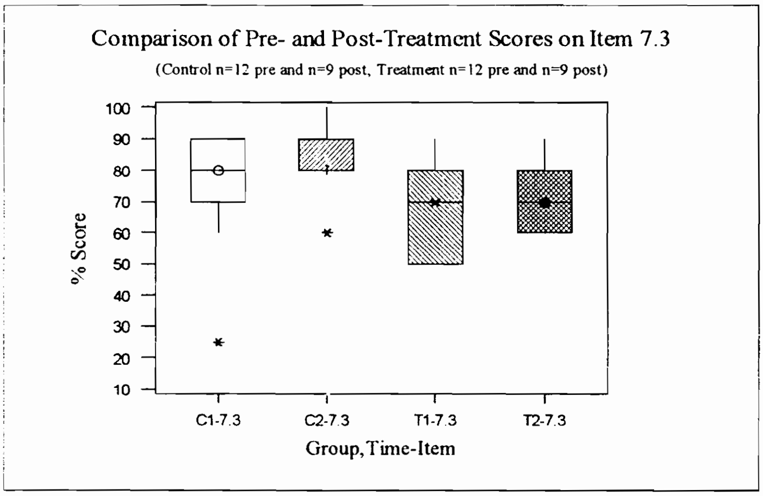
(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)

figure continues



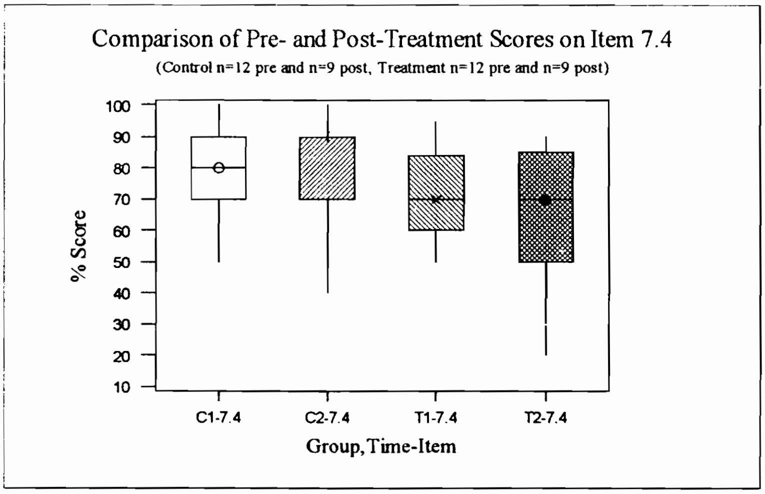
Comparison of Pre- and Post-Treatment Scores on Item 7.2

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 7.3

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)



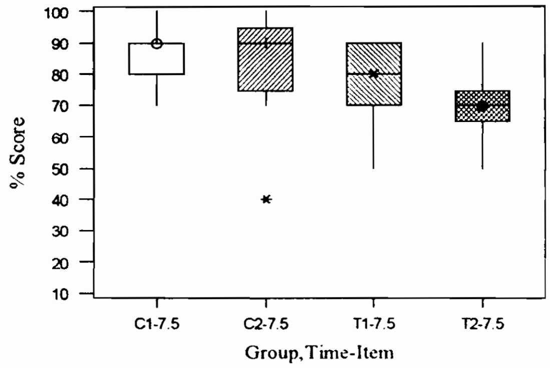
Comparison of Pre- and Post-Treatment Scores on Item 7.4

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)

figure continues

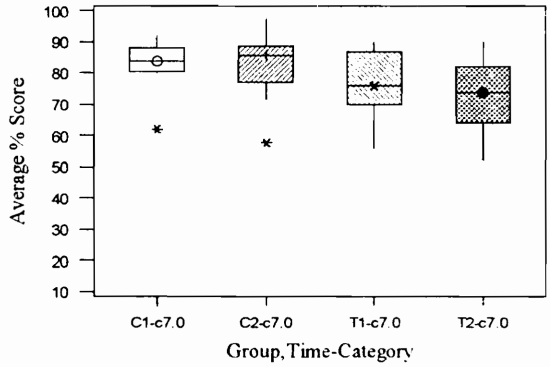
Comparison of Pre- and Post-Treatment Scores on Item 7.5

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 7.0

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)



Question 6 - Did agreement among evaluators on the score of an item change (improve) due to evaluator training?

Question six (Q6) addressed the comparison of the treatment group's scores from the initial evaluation (pre-training) to their scores from the second evaluation (post-training). The control group's first and second evaluation scores were also examined to account for any learning effect. Specifically, Q6 compared the relative agreement (i.e., dispersion of scores) within each group during their first and second evaluations. Q6 was addressed by visually examining the relative agreement displayed by the evaluators on the score of each item and by testing hypothesis 7 (H7). The relative agreement of the treatment group before and after training can be seen in the box plots of the item scores in Figure 7. The relative agreement of the control group for their first versus second evaluation can also be seen in Figure 7.

H7: Item score variances will be smaller for second evaluation scores than for first evaluation scores.

H7 tested for differences in item score variance between the first and second evaluations for each group. A related hypothesis, H7b, tested for differences between the treatment and control groups' second evaluations (a second evaluation version of H2).

> H7b: Item score variances will be smaller for the treatment group's second evaluation scores than for the control group's second evaluation scores.

Since the training was expected to increase agreement on scores, the treatment group was expected to exhibit less score variation in their second evaluation scores. The control group was expected to show little or no difference in score variation from the first to second evaluation. The treatment group was expected to show less variation than the control group on their second evaluation scores. H7 was tested using the screened first and second evaluation scores from H6. The specific hypotheses tested for H7 are shown below.

Testing Hypothesis 7

F-tests were used to compare the item score variances between the first and second evaluations for each group and between the treatment and control groups' second evaluations. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 6). A test-wise Type I error rate of alpha = 0.05 was used.

Tested Hypotheses

The following hypotheses were tested for each item. For the treatment group:

H_0: The variance of the treatment group's first evaluation scores on Item X.X is less than or equal to the variance of the treatment group's second evaluation scores on Item X.X.
H_1: The variance of the treatment group's first evaluation scores on Item X.X is GREATER than the variance of the treatment group's second evaluation scores on Item X.X.
where X.X = 1.1 to 7.5.

For the control group:

H_0: The variance of the control group's first evaluation scores on Item X.X is less than or equal to the variance of the control group's second evaluation scores on Item X.X.
H_1: The variance of the control group's first evaluation scores on Item X.X is GREATER than the variance of the control group's second evaluation scores on Item X.X.
where X.X = 1.1 to 7.5.

For the second evaluation (post-training) comparison:

H_0: The variance of the control group's second evaluation scores on Item X.X is less than or equal to the variance of the treatment group's second evaluation scores on Item X.X.
H_1: The variance of the control group's second evaluation scores on Item X.X is GREATER than the variance of the treatment group's second evaluation scores on Item X.X.
where X.X = 1.1 to 7.5.

Test Results

Tables 13 and 14 provide summaries of the F-test results by group and by item. Each table summarizes the results of comparing the first and second evaluation score variances for that group. Table 15 provides a summary of the F-test results for the comparison of the control and treatment groups' second evaluation item score variances.

Table 13.

Summary of Treatment Group Test Results for Hypothesis 7.

test of first and second evaluation score variances for Item	df_{num}, df_{den}	F_{obs}	p-value
1.1	12,10	0.433	0.9143
1.2	12,10	0.793	0.6532
1.3	12,10	0.538	0.8463
2.1	10,8	0.485	0.8593
2.2	10,8	2.738	0.0834
2.3	10,8	2.703	0.0861
3.1	12,8	1.292	0.3671
3.2	12,8	0.523	0.8498
4.1	13,9	1.160	0.4216
4.2	13,9	1.107	0.4507
4.3	13,9	2.654	0.0737
4.4	13,9	1.867	0.1756
5.1	10,11	1.637	0.2154
5.2	10,11	0.884	0.5733
5.3	10,11	1.233	0.3665
5.4	10,11	0.342	0.9490
6.1	11,8	0.341	0.9492
6.2	11,8	0.938	0.5517
6.3	11,8	0.513	0.8491
7.1	11,8	0.444	0.8940
7.2	11,8	1.417	0.3172
7.3	11,8	2.135	0.1455
7.4	11,8	0.384	0.9282
7.5	11,8	1.344	0.3449

Table 14.

Summary of Control Group Test Results for Hypothesis 7.

test of first and second evaluation variances for Item	df_{num}, df_{den}	F_{obs}	p-value
1.1	11,10	1.823	0.1767
1.2	11,10	0.956	0.5323
1.3	11,10	5.463	0.0061**
2.1	11,6	0.600	0.7810
2.2	11,6	0.893	0.5891
2.3	11,6	0.472	0.8670
3.1	10,9	0.413	0.9077
3.2	10,9	0.861	0.5933
4.1	10,8	6.665	0.0065**
4.2	10,8	2.774	0.0808
4.3	10,8	0.826	0.6191
4.4	10,8	6.743	0.0062**
5.1	8,9	1.154	0.4144
5.2	8,9	1.051	0.4665
5.3	8,9	1.214	0.3867
5.4	8,9	1.364	0.3254
6.1	6,11	0.534	0.7722
6.2	6,11	2.100	0.1356
6.3	6,11	1.525	0.2573
7.1	11,8	1.189	0.4129
7.2	11,8	0.688	0.7234
7.3	11,8	2.888	0.0715
7.4	11,8	0.551	0.8227
7.5	11,8	0.283	0.9718

** Significant at the 0.01 level (test-wise error rate was 0.05).

Table 15.

Summary of Control versus Treatment Group (2nd evaluation scores) Test Results for Hypothesis 7b.

test of control and treatment group second evaluation score variances for Item	df_{num}, df_{den}	F_{obs}	p-value
1.1	10,10	0.301	0.9642
1.2	10,10	0.861	0.5912
1.3	10,10	0.301	0.9642
2.1	6,8	2.181	0.1520
2.2	6,8	1.677	0.2437
2.3	6,8	2.865	0.0853
3.1	9,8	2.294	0.1282
3.2	9,8	0.992	0.5100
4.1	8,9	0.116	0.9971
4.2	8,9	0.506	0.8249
4.3	8,9	0.891	0.5593
4.4	8,9	0.552	0.7926
5.1	9,11	2.560	0.0721
5.2	9,11	1.466	0.2708
5.3	9,11	1.222	0.3708
5.4	9,11	0.934	0.5331
6.1	11,8	0.674	0.7337
6.2	11,8	0.548	0.8248
6.3	11,8	0.428	0.9036
7.1	8,8	0.586	0.7668
7.2	8,8	1.641	0.2496
7.3	8,8	1.125	0.4359
7.4	8,8	0.667	0.7100
7.5	8,8	2.801	0.0833

Question 7 - Did agreement among evaluators on the score of a category change (improve) due to evaluator training?

Question seven (Q7) compared the treatment group's category scores from the initial evaluation (pre-training) to their category scores from the second evaluation (post-training). The control group's first and second evaluation category scores were also examined to account for any learning effect. Like Q6, Q7 examined the relative agreement (i.e., dispersion of scores) within each group during their first and second evaluations; however, Q7 focused on category rather than item scores. Q7 was addressed by visually examining the relative agreement displayed by the evaluators on the score of each category and by testing hypothesis 8 (H8). The relative agreement of the treatment group before and after training can be seen in the box plots of the category scores in Figure 7. The relative agreement of the control group for their first versus second evaluation can also be seen in Figure 7.

H8: Category score variances will be smaller for second evaluation scores than for first evaluation scores.

H8 tested for differences in category score variance between the first and second evaluations for each group. A related hypothesis, H8b, tested for differences in category score variance between the treatment and control groups' second evaluations (a second evaluation version of H3).

> H8b: Category score variances will be smaller for the treatment group's second evaluation scores than for the control group's second evaluation scores.

Since the training was expected to increase agreement on scores, the treatment group was expected to exhibit less score variation in their second evaluation scores. The control group was expected to show little or no difference in score variation from the first to second evaluation. The treatment group was expected to show less variation than the control group on their second evaluation scores. H8 was tested using the screened first and second evaluation scores from H6. The specific hypotheses tested for H8 are shown below.

Testing Hypothesis 8

F-tests were used to compare the category score variances between the first and second evaluations for each group and between the treatment and control groups' second evaluations. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 6). A test-wise Type I error rate of alpha = 0.05 was used.

Tested Hypotheses

The following hypotheses were tested for each category. For the treatment group:

H_0: The variance of the treatment group's first evaluation scores on Category X.0 is less than or equal to the variance of the treatment group's second evaluation scores on Category X.0.
H_1: The variance of the treatment group's first evaluation scores on Category X.0 is GREATER than the variance of the treatment group's second evaluation scores on Category X.0.
where X.0 = 1.0 to 7.0.

For the control group:

H_0: The variance of the control group's first evaluation scores on Category X.0 is less than or equal to the variance of the control group's second evaluation scores on Category X.0.
H_1: The variance of the control group's first evaluation scores on Category X.0 is GREATER than the variance of the control group's second evaluation scores on Category X.0.
where X.0 = 1.0 to 7.0.

For the second evaluation (post-training) comparison:

H_0: The variance of the control group's second evaluation scores on Category X.0 is less than or equal to the variance of the treatment group's second evaluation scores on Category X.0.
H_1: The variance of the control group's second evaluation scores on Category X.0 is GREATER than the variance of the treatment group's second evaluation scores on Category X.0.
where X.0 = 1.0 to 7.0.

Test Results

Tables 16 and 17 provide summaries of the F-test results by group and by item. Each table summarizes the results of comparing the first and second evaluation score variances for that group. Table 18 provides a summary of the F-test results for the comparison of the control and treatment groups' second evaluation item score variances.

Table 16.

Summary of Treatment Group Test Results for Hypothesis 8.

test of variances for Category	df_{num}, df_{den}	F_{obs}	p-value
1.0	12,10	0.546	0.8406
2.0	10,8	1.563	0.2694
3.0	12,8	0.812	0.6406
4.0	13,9	1.794	0.1916
5.0	10,11	1.100	0.4362
6.0	11,8	0.365	0.9379
7.0	11,8	0.797	0.6454

Table 17.

Summary of Control Group Test Results for Hypothesis 8.

test of variances for Category	df_{num}, df_{den}	F_{obs}	p-valuc
1.0	11,10	3.905	0.0202*
2.0	11,6	0.532	0.8275
3.0	10,9	0.572	0.8017
4.0	10,8	5.207	0.0141*
5.0	8,9	0.972	0.5105
6.0	6,11	0.721	0.6417
7.0	11,8	0.446	0.8927

* = significant at the 0.05 level

Table 18.

Summary of Control versus Treatment Group (2nd evaluation scores) Test Results for Hypothesis 8b.

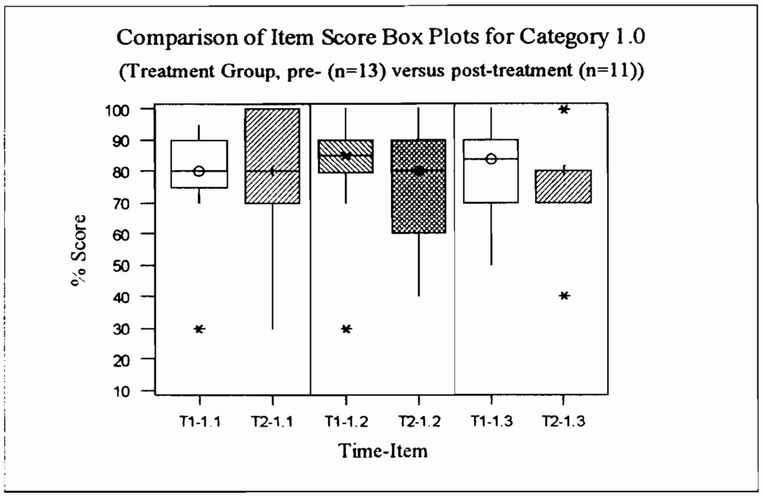
test of control and treatment group second evaluation score variances for Category	df_{num}, df_{den}	F_{obs}	p-value
1.0	10,10	0.217	0.988
2.0	6,8	2.263	0.1413
3.0	9,8	1.713	0.2300
4.0	8,9	0.233	0.9739
5.0	9,11	1.614	0.2241
6.0	11,8	0.544	0.8276
7.0	8,8	0.910	0.5514

Question 8 - Did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?

Question 8 (Q8) examined the consistency of score dispersion between the items of a category and was examined for each category. Q8 was addressed by comparing the relative variation (dispersion) of item scores across all the items of a category and by further analyzing the summary results from H7. Because the variance of item scores within a category were not independent, a testable hypothesis was not developed to address Q8. To address the issue of change due to training, the pre- and post-treatment scores of the treatment group were graphically compared. Box plots for all the items of a category were plotted on a single chart to facilitate this comparison (see Figure 8). To account for learning effect over time, the same comparison can be made for the control group by comparing item score box plots across the items of each category in Figure 7. The data used for these box plots were the screened data from H6. The items within a category measure related constructs; therefore, little or no difference was expected between the dispersions of item scores across a particular category (for a given evaluation). The dispersion of item scores was expected to decrease from the first to the second evaluation.

H7 tested for differences in item score variance over time on an item-by-item basis. The results of H7 were analyzed to determine the number of items for which the score variance improved (decreased) for each group. For the treatment group, the variance increased for twelve items and the variance decreased for twelve items. None of these changes were statistically significant. For the control group, the variance increased for eleven items and the variance decreased for thirteen items. Three of these changes were statistically significant. In each case the change represented a decrease in variance from the first to the second evaluation and the significance was strong ($p \leq 0.01$).

Category 1.0 - Leadership



Category 2.0 - Information and Analysis

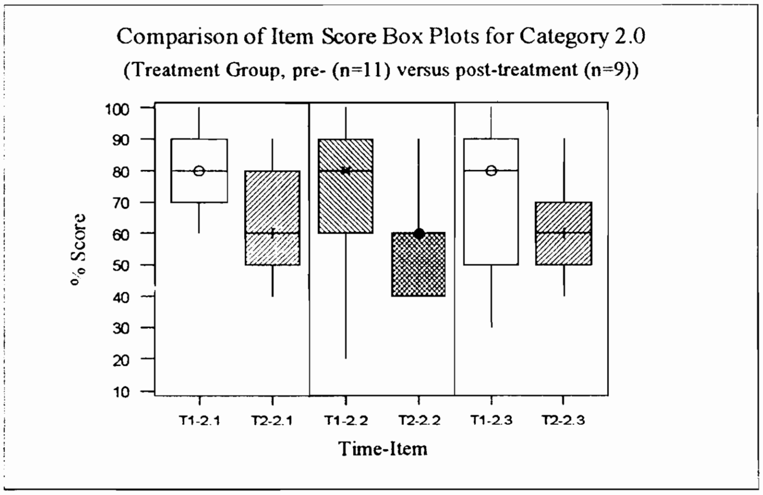
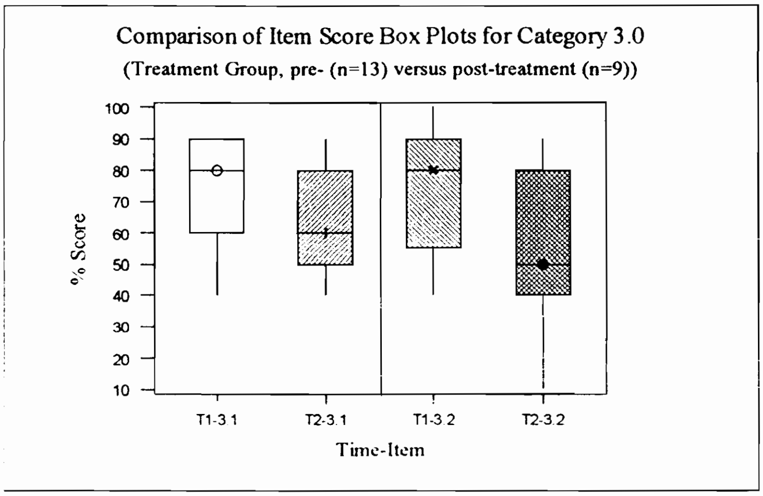


Figure 8. Box plot comparisons of item scores across each category (pre- versus post-treatment)

figure continues

Category 3.0 - Strategic Planning


Comparison of Item Score Box Plots for Category 3.0
(Treatment Group, pre- (n=13) versus post-treatment (n=9))

Category 4.0 - Human Resource Development and Management

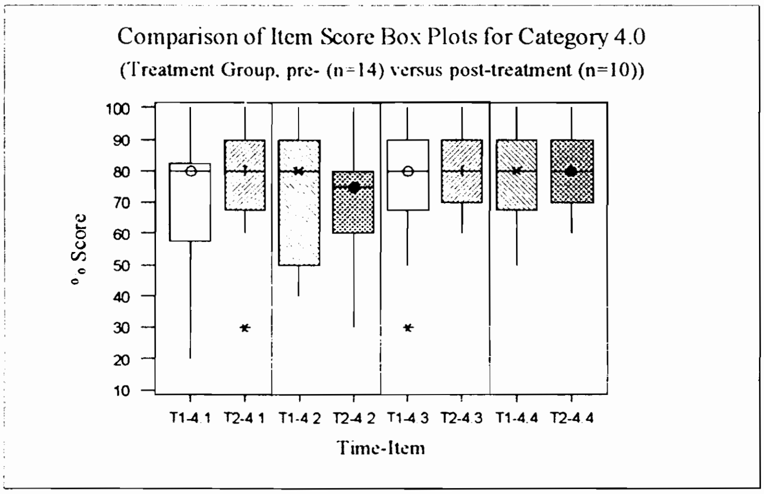
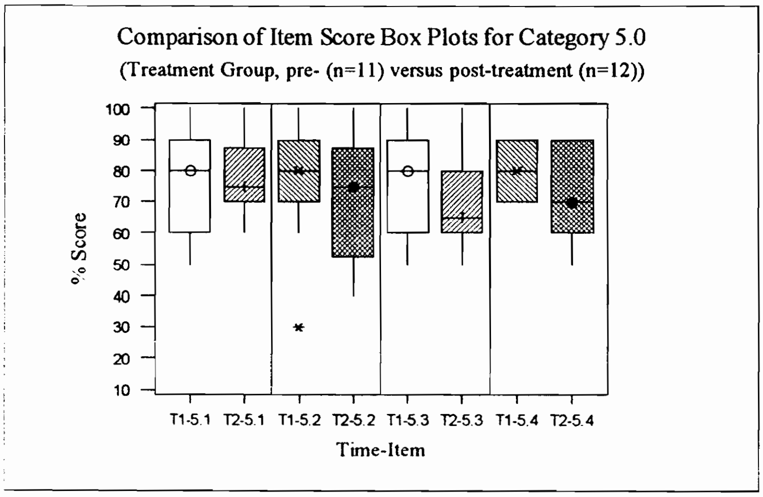

Comparison of Item Score Box Plots for Category 4.0
(Treatment Group, pre- (n=14) versus post-treatment (n=10))

figure continues

Category 5.0 - Process Management



Category 6.0 - Business Results

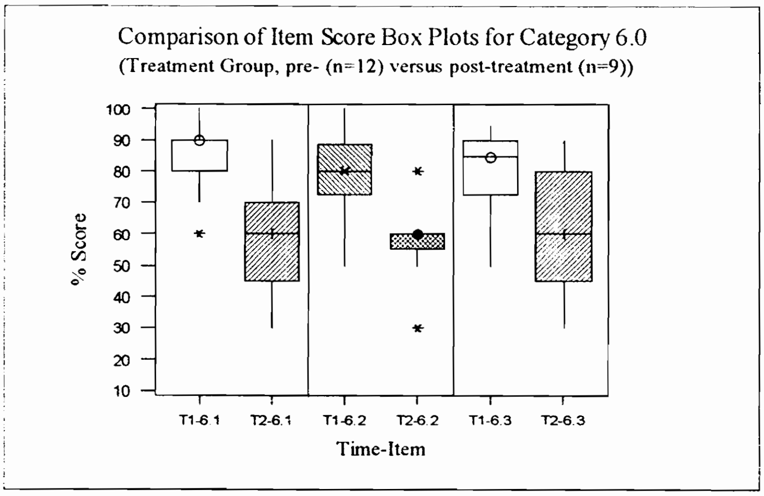
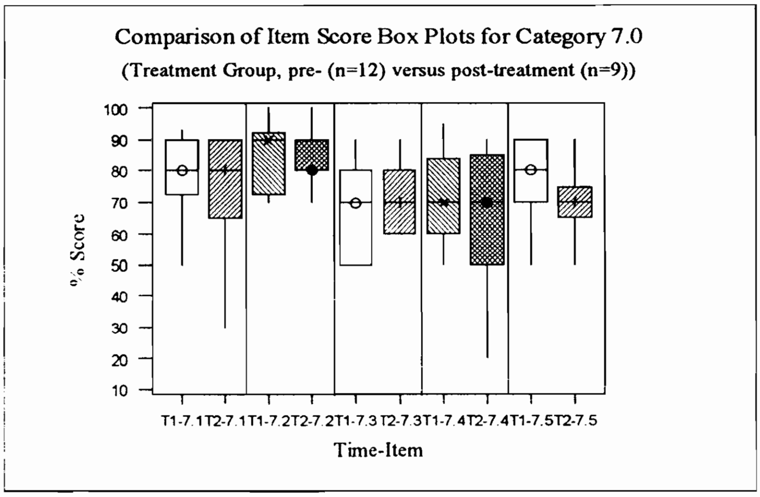


figure continues

Category 7.0 - Customer Focus and Satisfaction



Question 9 - Did within-category variation of the evaluator scores across all seven categories change (decrease) due to evaluator training?

Question 9 (Q9) examined the consistency of category score dispersion across all seven categories both before and after training. Q9 was addressed by graphically comparing the relative dispersion of category scores across all seven categories, testing hypothesis 9 (H9), and reviewing the results of the analysis of Q4. Hypothesis 9 tested for differences in category score variances between the seven categories and was tested for each group. Q4 was the pre-treatment examination of category score dispersion across all seven categories.

H9: There will be a difference in score variances between categories for both the control and treatment groups.

Since statistical differences were found between the control group's pre- and post-treatment category score variances[16], the group scores were not pooled for this analysis. H9 was tested using the screened control and treatment groups' scores from H6. The specific hypotheses tested to address H9 are shown below. The relative dispersion of category scores across all seven categories can be seen in the box plots of Figures 9 and 10.

Since each category measures a different construct, a difference was expected to be seen between the score variances of the seven categories. This difference was seen when H4 tested the category score variances of the combined groups' pre-treatment scores. If training has a significant effect, the inherent difference in score variances between categories might be reduced below statistical significance for the treatment group's second evaluation. The difference between score variances across the seven categories was expected to be seen in the control group's second evaluation.

[16] No significant differences were found between the treatment group's pre- and post-treatment category score variances. Also, combined post-treatment scores would likely have larger variances than combined pre-treatment scores due to the downward shift in the means of the treatment group's post-treatment scores.

Testing Hypothesis 9

Hartley's F_{max} test (Ott, 1984) was used to test for homogeneity of category variances. Prior to analysis, the data were screened to remove responses suspected of contamination (described under Hypothesis 6). A Type I error rate of alpha = 0.05 was used. Since the test for differences in score variances between categories was significant for the control group, pairwise comparisons were conducted to provide additional evidence regarding differences in the control group's category score variances.

Tested Hypotheses

The following hypothesis was tested using the second evaluation scores of the treatment group.

H_0: There is no difference in the variances of trained evaluators' second evaluation category scores across the seven categories.

H_1: There is a difference in the variances of trained evaluators' second evaluation category scores across the seven categories.

The following hypothesis was tested using the second evaluation scores of the control group.

H_0: There is no difference in the variances of untrained evaluators' second evaluation category scores across the seven categories.

H_1: There is a difference in the variances of untrained evaluators' second evaluation category scores across the seven categories.

The test statistic and rejection region for H9 are shown in Appendix AW.

Test Results

Figures 9 and 10 display the box plots of the second evaluation category scores for the treatment and control groups, respectively. Since Q4 was addressed using combined group data, separate boxplots of the first evaluation category scores for the treatment and control groups were constructed for comparison purposes (see Figures 11 and 12). Table 19 lists the standard deviations and resulting variances of the second evaluation scores for each category for both the treatment and control groups. The maximum and minimum variances from Table 19 were used to produce the F_{max} statistics below.

$F_{max\ obs.} = 2.69$ (treatment group)

$F_{max\ obs.} = 15.06^{**}$ (control group)

** = significant at the 0.01 level, $F_{max(7,\ 9)0.99} = 13.1$

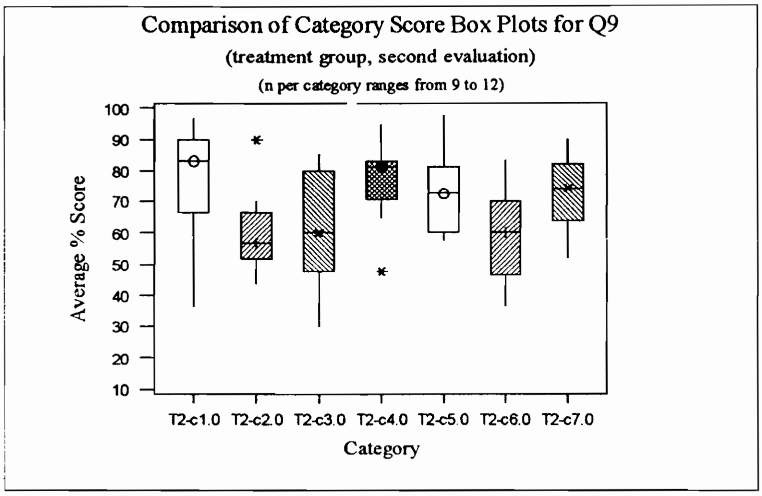


Figure 9. Box plot comparisons of the treatment group's second evaluation category scores.

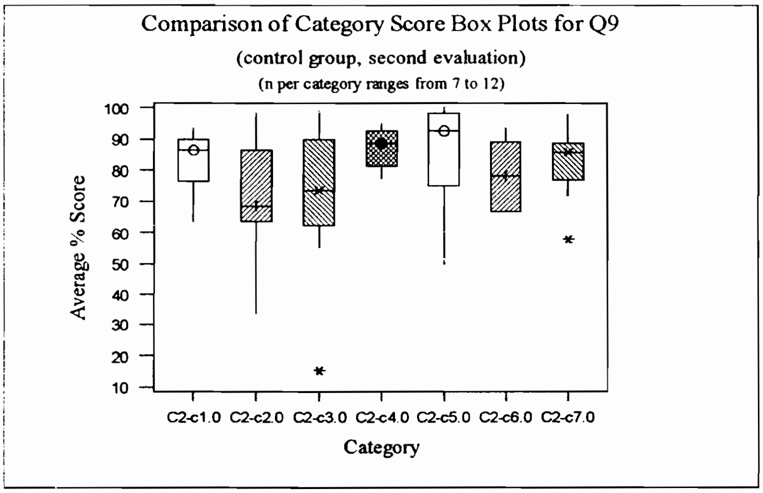


Figure 10. Box plot comparisons of the control group's second evaluation category scores.

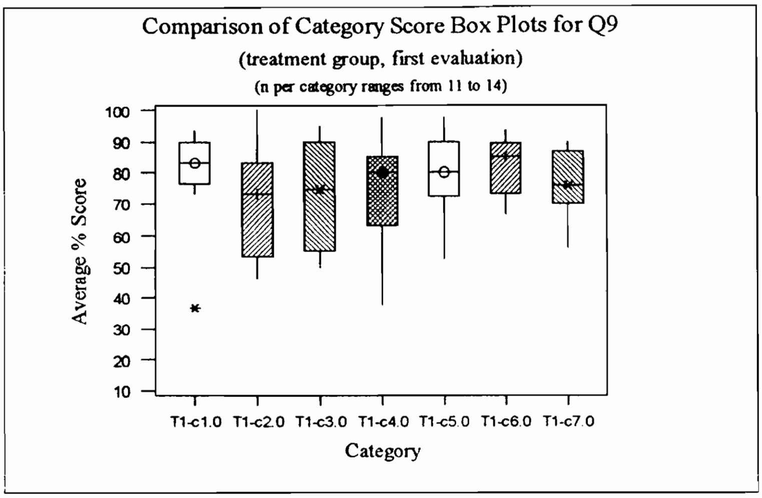


Figure 11. Box plot comparisons of the treatment group's first evaluation category scores.

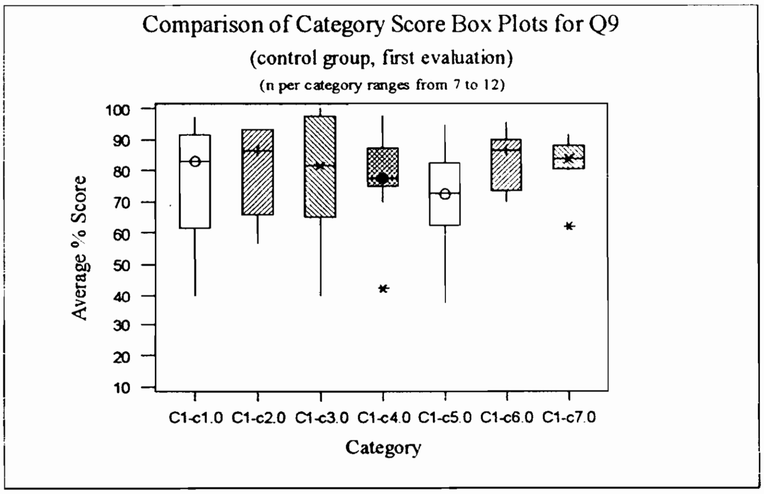


Figure 12. Box plot comparisons of the control group's first evaluation category scores.

Table 19.

Standard Deviations and Variances of the Scores of each Category (second evaluation)

Category	Std. Dev.	Variance	F-obs	n
T2-c1.0	19.82	392.83	numerator	11
T2-c2.0	13.52	182.79		9
T2-c3.0	18.62	346.70		9
T2-c4.0	13.02	169.52		10
T2-c5.0	13.05	170.30		12
T2-c6.0	15.16	229.83		9
T2-c7.0	12.09	146.17	denominator	9
		Fmax =	2.69	
		n average =	9.86	df = 9
C2-c1.0	9.24	85.38		11
C2-c2.0	20.34	413.72		7
C2-c3.0	24.37	593.90	numerator	10
C2-c4.0	6.28	39.44	denominator	9
C2-c5.0	16.58	274.90		10
C2-c6.0	11.18	124.99		12
C2-c7.0	11.53	132.94		9
		Fmax =	15.06	
		n average =	9.71	df = 9

Pairwise Comparisons (H9)

Pairwise comparisons were conducted to determine which of the untrained group's second evaluation category score variances were different. The variance of the scores of each category was compared to the variance of the scores of each of the other six categories. The data used for these comparisons were the screened data from Hypothesis 6, further screened to reduce possible dependencies. Since each subject evaluated two categories, using scores from the same subject for both categories of a comparison violates the assumption of independence. To compensate for this, the second evaluation data were screened as described below.

Data Screening to Eliminate Dependencies

The following procedure was used for each paired comparison. The data for all the control group subjects who evaluated either of the two categories being compared were compiled into a spreadsheet. Those subjects evaluating both categories were numbered 1 to N. The next step was to assign each of these subjects to one of two groups, each group N/2 in size. Microsoft Excel's random number generator was used to produce a table of random numbers with a uniform distribution and a range greater than or equal to 1 to N. A new column of random numbers was used for each comparison, starting alternately at the top or bottom of the column. Each random number was used to identify which of the 1 to N subjects were assigned to the first group (i.e., their scores from the second category would be discarded for this analysis). Those remaining after N/2 had been assigned to the first group were assigned to the second group (i.e., their scores from the first category would be discarded for this analysis). When N was an odd number, a coin flip was used to determine which group would receive the additional subject. The result was elimination of the dependency while maintaining sample size as much as possible. Subsequent comparisons started with the complete data set and followed the same screening procedure.

Tested Hypotheses

F-tests were used to compare the score variances of each pair of categories. An experiment-wise Type I error rate of alpha = 0.05 was used. With the Bonferroni adjustment, this resulted in a comparison-wise error rate of alpha = 0.0024. See Appendix AH for the master spreadsheet and an example spreadsheet used to calculate the variances for the pairwise comparisons.

H_0: There is no difference between the variance of untrained evaluators' second evaluation Category X.0 scores and the variance of untrained evaluators' second evaluation Category Y.0 scores.

H_1: There is a difference between the variance of untrained evaluators' second evaluation Category X.0 scores and the variance of untrained evaluators' second evaluation Category Y.0 scores.

where X.0 = 1.0 to 6.0 and Y.0 = 2.0 to 7.0 (no category was compared against itself).

Test Results

Table 20 lists the statistics and results for each comparison. Table 21 provides a summary of the p-values in matrix form to facilitate evaluation. That is, Table 21 provides the results in a form that highlights patterns across the pairwise comparisons.

Table 20.

Test Results for H9 Control Group Pairwise Comparisons.

Comparison of variances for Category X.0 to Y.0	Variance X.0, Variance Y.0	df_{num}, df_{den}	F_{obs}	$P(F \leq F_{obs})$ or (1-p/2)	p-value
1.0 to 2.0	85.5, 413.6	6,10	4.837	0.9856	0.0288*
1.0 to 3.0	85.5, 621.4	8,10	7.268	0.9974	0.0052*
1.0 to 4.0	93.3, 43.5	9,7	2.145	0.8366	0.3268
1.0 to 5.0	85.5, 274.9	9,10	3.215	0.9585	0.0830
1.0 to 6.0	94.4, 119.4	8,7	1.265	0.6152	0.7696
1.0 to 7.0	89.0, 150.2	7,9	1.688	0.7721	0.4558
2.0 to 3.0	302.7, 629.6	8,5	2.080	0.7821	0.4358
2.0 to 4.0	496.0, 31.2	5,7	15.897	0.9989	0.0022**
2.0 to 5.0	190.7, 282.1	8,5	1.479	0.6538	0.6924
2.0 to 6.0	413.6, 125.0	6,11	3.309	0.9589	0.0822
2.0 to 7.0	496.3, 133.0	5,8	3.732	0.9515	0.0970
3.0 to 4.0	593.9, 42.2	9,7	14.073	0.9989	0.0022**
3.0 to 5.0	668.0, 283.0	8,7	2.360	0.8626	0.2748
3.0 to 6.0	662.0, 125.0	8,11	5.296	0.9934	0.0132*
3.0 to 7.0	222.4, 150.2	8,7	1.481	0.6910	0.6180
4.0 to 5.0	39.9, 274.9	9,7	6.890	0.9907	0.0186*
4.0 to 6.0	34.3, 125.0	11,7	3.644	0.9514	0.0972
4.0 to 7.0	37.7, 147.4	7,7	3.910	0.9537	0.0926
5.0 to 6.0	317.2, 129.6	7,10	2.447	0.9033	0.1934
5.0 to 7.0	274.9, 114.2	9,7	2.407	0.8701	0.2598
6.0 to 7.0	123.0, 133.0	8,10	1.081	0.5550	0.8900

* Significant at the 0.05 level (experiment-wise error rate).

** Significant at the 0.0024 level (comparison-wise error rate with the Bonferroni adjustment).

Table 21.

P-Values for Each H9 Control Group Pairwise Comparison

Category	2.0	3.0	4.0	5.0	6.0	7.0
1.0	0.0288*	0.0052*	0.3268	0.0830	0.7696	0.4558
2.0		0.4358	0.0022**	0.6924	0.0822	0.0970
3.0			0.0022**	0.2748	0.0132*	0.6180
4.0				0.0186*	0.0972	0.0926
5.0					0.1934	0.2598
6.0						0.8900

* Significant at the 0.05 level (experiment-wise error rate).

** Significant at the 0.0024 level (comparison-wise error rate with the Bonferroni adjustment).

Question 10 - Did the accuracy of the evaluators' scores change (improve) due to evaluator training?

Question 10 (Q10) examined the accuracy of the scores from a sample of evaluators before and after training. Q10 was addressed by graphical analysis, by testing hypothesis 10 (H10), and by conducting a content analysis of the evaluators' comments accompanying the score of each item. Expert scores[17] for each item and category were plotted onto the box plots from H6 to facilitate comparison with the distribution of each group's pre- and post-treatment scores (see Figure 13). The box plots of Figure 13 only illustrate scores, not direct measures of accuracy. To illustrate the distributions of each group's accuracy indices over time, group x time box plots were constructed for each accuracy index by category (see Appendix AI). Hypothesis 10 is the post-treatment version of H5, which tested for differences in pre-training score accuracy between the control and treatment group. H10 tested for differences in score accuracy between the first and second evaluations of both the control group and treatment group. The accuracy of each group was expected to improve due to learning over time. Improvement is reflected by a decrease in an accuracy index; perfect accuracy results in an index value of zero. Since the treatment group received training prior to completing their second evaluations, a statistically significant difference (improvement) in accuracy was expected for the treatment group over time. The change in the control group's accuracy over time was not expected to be statistically significant. The treatment group was also expected to display significantly better second evaluation score accuracy than the control group. The results of the content analysis of the evaluators' comments accompanying the score of each item were expected to support the results of testing H10.

H10: The accuracy of evaluators' scores will improve between the first and second evaluations.

[17] Expert scores were provided for each item as a ten point range. No information was provided regarding the exact number of experts involved or the actual distribution of the scores. For statistical purposes and graphical simplicity, the mean of this ten point range was used to estimate the true score.

H10 was tested using accuracy statistics calculated from the control group's and treatment group's first and second evaluation scores. Elevation[18] and dimensional accuracy[19] were calculated for each evaluator for each category. The elevation and DA indices for the first evaluation scores were calculated for H5 and used again here (see Table 8). The mean and median elevation and DA indices for the second evaluation scores are shown in Table 22. Procedures for calculating these statistics are described in

Table 22.

Group Mean and Median Accuracy Indices for each Category (2nd Evaluation)

Category	Accuracy Index	Control Mean	Treatment Mean	Control Median	Treatment Median
1.0	Elevation	24.71	24.39	28.30	25.00
2.0	Elevation	18.33	9.46	11.70	5.00
3.0	Elevation	36.65	22.78	33.75	20.00
4.0	Elevation	23.10	16.30	25.00	17.55
5.0	Elevation	31.04	16.72	36.30	16.30
6.0	Elevation	23.19	12.03	23.30	8.30
7.0	Elevation	22.33	14.56	25.00	13.00
1.0	DA	11.65	10.05	10.80	8.50
2.0	DA	10.97	9.50	8.50	8.50
3.0	DA	5.85	8.89	5.00	10.00
4.0	DA	10.49	11.42	9.80	10.40
5.0	DA	16.47	16.58	15.75	15.15
6.0	DA	6.54	6.73	6.20	4.10
7.0	DA	10.28	11.51	9.70	12.40

[18] Elevation is the average of J's predictions over all items and ratees minus the central tendency of the self-descriptions (i.e., true scores) for all items and ratees combined (Cronbach, 1955, p. 178). Sulsky and Balzer's (1988) two-way ANOVA analogy described Elevation as the difference between the rater's grand mean and the true score grand mean.

[19] Dimensional accuracy (DA) is "equivalent to Cronbach's (1955) differential accuracy in that it measures the accuracy with which each rater evaluated a single ratee on each dimension" (Hauenstein and Alexander, 1991, p. 308). Dimensional accuracy is similar to McIntyre, Smith, and Hassett's correlational accuracy, in that it "measures the parallelism between subjects' scores and true scores" (1984, p. 151)

Appendix AT. The group x time elevation and DA indices for each category were tested for normality using the Kolmogorov-Smirnov test (Coakley, 1995; Gibbons, 1985). Since several of the indices rejected the hypothesis of normality from the K-S test, both classical and rank-based tests were performed for each index and category. If the results of these tests were consistent, then the classical procedures were assumed to be robust enough to handle the violation of the normality assumption.

Testing Hypothesis 10

A two-factor ANOVA was used to test for an interaction between group and time on each accuracy index for each category. If no interaction effect was seen, group (row) and time (column) effects were examined. If mild ($p \leq 0.10$) or stronger evidence[20] of an interaction was present, simple group and time effects were tested. When there was little evidence of an interaction effect, but mild ($p \leq 0.10$) or stronger evidence of a main effect, the simple effects related to that main effect were tested. T-tests were used for all simple effects tests. Although Minitab was used to calculate actual p-values, a test-wise Type I error rate of alpha = 0.05 was used to identify significant results. Prior to analysis, the data were screened as described under H1 and H6. The specific hypotheses tested to address H10 are shown below. A content analysis of the evaluator's comments[21] on strengths, areas for improvement, and site visit issues was used to compare the evaluator's comments to the expert's comments.

The following rank-based tests were used to confirm the results of the classic tests described above. A Friedman-type rank test (Mack and Skillings, 1980) was used to test for the main effects in the two-factor ANOVA. A Mann-Whitney test (Gibbons, 1985) was used to test for simple effects. Since the rank-based procedures were unable to test

[20] This lower standard of evidence ($p \leq 0.10$) is only used to determine if further analysis is warranted, not to determine statistical significance.

[21] These comments accompanied the score of each item and were intended to provide justification for the score.

for interaction effects, the simple effects related to any observed interaction effect were used to compare the results of the classical and rank-based tests.

Tested Hypotheses

The group x time ANOVA tested the following hypotheses for each accuracy index and each category. The Friedman-type rank test tested similar hypotheses, with the omission of the hypothesis regarding interaction effects.

H_0: There are no interaction effects when comparing evaluators' mean accuracy indices based on both group (untrained versus trained) and time (first evaluation versus second evaluation).
H_0: There are no group effects when comparing evaluators' mean accuracy indices averaged over time (first evaluation and second evaluation).
H_0: There are no time effects when comparing evaluators' mean accuracy indices averaged across groups (untrained and trained).
H_1: There is an effect (interaction, group, or time).
Accuracy is measured using elevation for the first series of tests, then dimensional accuracy for the second series of tests.

The following hypothesis was tested using a t-test when simple group effects were examined. These tests were conducted using second evaluation (post-treatment) data, since the same hypothesis was already tested for H5 using first evaluation (pre-treatment) data. The corresponding hypotheses, tested with the Mann-Whitney procedure, tested the medians rather than the means.

H_0: The mean accuracy index of the untrained evaluators' second evaluation scores on Category X.0 is less than or equal to the mean accuracy index of the trained evaluators' second evaluation scores on Category X.0.
H_1: The mean accuracy index of the untrained evaluators' second evaluation scores on Category X.0 is GREATER THAN the mean accuracy index of the trained evaluators' second evaluation scores on Category X.0.
Where X.0 may be any category from 1.0 to 7.0; and a smaller accuracy index represents greater (improved) accuracy for both elevation and DA.

The following hypotheses were tested using a t-test when simple time effects were examined. The first pair of hypotheses applied to the untrained evaluators and the second pair to the trained evaluators. The related hypotheses, tested with the Mann-Whitney procedure, tested the medians rather than the means.

H_0: The mean accuracy index of the untrained evaluators' first evaluation scores on X.0 is less than or equal to the mean accuracy index of the untrained evaluators' second evaluation scores on X.0.

H_1: The mean accuracy index of the untrained evaluators' first evaluation scores on X.0 is GREATER THAN the mean accuracy index of the untrained evaluators' second evaluation scores on X.0.
Where X.0 may be any category from 1.0 to 7.0; and a smaller accuracy index represents greater (improved) accuracy for both elevation and DA.

H_0: The mean accuracy index of the trained evaluators' first evaluation scores on X.0 is less than or equal to the mean accuracy index of the trained evaluators' second evaluation scores on X.0.

H_1: The mean accuracy index of the trained evaluators' first evaluation scores on X.0 is GREATER THAN the mean accuracy index of the trained evaluators' second evaluation scores on X.0.
Where X.0 may be any category from 1.0 to 7.0; and a smaller accuracy index represents greater (improved) accuracy for both elevation and DA.

The test statistics and rejection regions for the classical and rank-based tests are described in Appendix AX.

Test Results

The results of the two-factor ANOVAs and the tests for simple effects are summarized in Table 23. Complete ANOVA tables for each accuracy index by category are shown in Appendix AJ: Edited Minitab Session Files of 2-Factor ANOVA Results for Hypothesis 10. The pre-treatment (first evaluation) simple group effects were tested for H5 and the results were shown in Table 9. Table 24, Summary of Friedman-Type Rank Test and Mann-Whitney Test Results for Hypothesis 10, provides the results of the equivalent rank-based tests to those shown in Table 23. Group x time graphs of mean accuracy and median accuracy (see Appendix AL) were constructed to assist in the interpretation of the results shown in Tables 23 and 24.

Table 23.

Summary of ANOVA and t-test Results for Hypothesis 10

Category	Interaction	Main Effects		Simple Effects			
- Accuracy	Effect	Group	Time	Group		Time	
Index				Pre-	Post-	Control	Treatment
1.0 - E	0.573	0.651	0.827	see Notes	0.47	0.68	0.38
2.0 - E	0.499	0.121	0.077*		0.067	0.25	0.023**
3.0 - E	0.438	0.058*	0.220		0.030**	0.38	0.08
4.0 - E	0.227	0.138	0.615		0.028**	0.90	0.31
5.0 - E	0.028**	~~0.172~~	~~0.436~~		0.011**	0.98	0.14
6.0 - E	0.164	0.047**	0.003**		0.013**	0.15	0.001**
7.0 - E	0.760	0.016**	0.806		0.045**	0.52	0.35
1.0 - DA	0.518	0.701	0.611		0.26	0.78	0.46
2.0 - DA	0.569	0.847	0.348		0.25	0.40	0.14
3.0 - DA	0.423	0.342	0.738		0.86	0.22	0.63
4.0 - DA	0.125	0.310	0.307		0.65	0.058	0.66
5.0 - DA	0.343	0.371	0.850		0.52	0.32	0.82
6.0 - DA	0.627	0.709	0.344		0.54	0.15	0.38
7.0 - DA	0.211	0.648	0.443		0.68	0.38	0.95

Notes:

* = mild evidence of an interaction or main effect, $p \leq 0.10$

** = evidence of significant effect, $p \leq 0.05$

The pre-treatment simple group effects were tested for H5 and the results are summarized in Table 9. Testing H5 found no significant simple group effects for either E or DA.

Table 24.

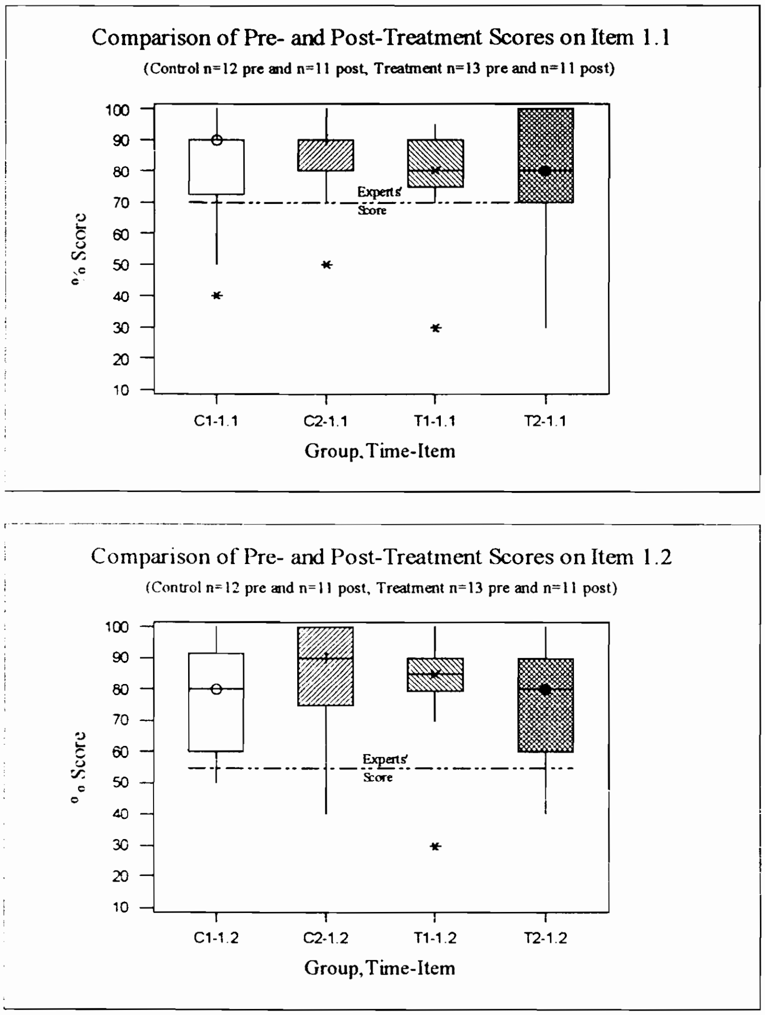
Summary of Friedman-Type Rank Test and Mann-Whitney Test Results for Hypothesis 10

Category - Accuracy Index		Main Effects		Simple Effects			
		Group	Time	Group		Time	
				Pre-	Post-	Control	Treatment
1.0 - E	na	0.8493	0.8751	see Notes	0.4735	>0.50	0.4305
2.0 - E	na	0.1027	0.0478**		0.0606	0.2895	0.0159**
3.0 - E	na	0.0481**	0.1712		0.0268**	0.2968	0.0894
4.0 - E	na	0.2423	0.6000		0.0493**	>0.50	0.2870
5.0 - E	na	0.2513	0.7081		0.0123**	>0.50	0.1263
6.0 - E	na	0.0569*	0.0042**		0.0160**	0.1251	0.0027**
7.0 - E	na	0.0260**	1.0000		0.0419**	>0.50	0.3867
1.0 - DA	na	0.9542	0.6010		0.3342	>0.50	0.4423
2.0 - DA	na	0.7079	0.6297		0.1829	>0.50	0.1517
3.0 - DA	na	0.4028	0.8515		>0.50	0.2578	>0.50
4.0 - DA	na	0.2761	0.2354		>0.50	0.0640	0.4416
5.0 - DA	na	0.3648	0.8501		0.4474	0.3717	>0.50
6.0 - DA	na	0.5056	0.3107		0.3855	0.1251	0.3998
7.0 - DA	na	0.5667	0.4662		>0.50	0.3879	>0.50

* = mild evidence of an interaction or main effect, $p \leq 0.10$
** = evidence of significant effect, $p \leq 0.05$

The main effects were tested with the modified Friedman-type rank test and the simple effects were tested with the Mann-Whitney procedure. The pre-treatment simple group effects were tested for H5 and the results are summarized in Table 9.

Category 1.0 - Leadership



Comparison of Pre- and Post-Treatment Scores on Item 1.1

(Control n=12 pre and n=11 post, Treatment n=13 pre and n=11 post)

Comparison of Pre- and Post-Treatment Scores on Item 1.2

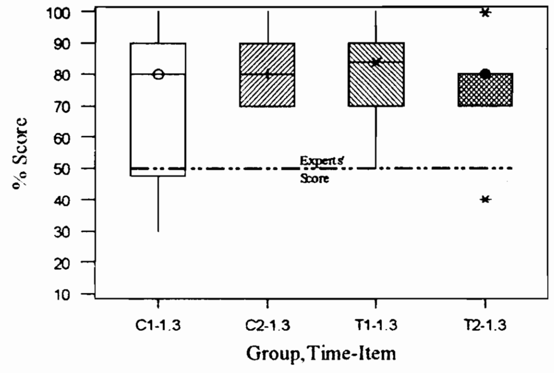
(Control n=12 pre and n=11 post, Treatment n=13 pre and n=11 post)

Figure 13. Comparison of expert scores to box plots of each group's pre- and post-treatment scores.

figure continues

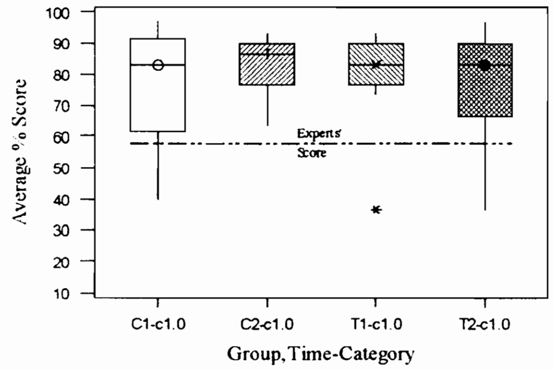
Comparison of Pre- and Post-Treatment Scores on Item 1.3

(Control n=12 pre and n=11 post, Treatment n=13 pre and n=11 post)



Comparison of Pre- and Post-Treatment Scores on Category 1.0

(Control r.=12 pre and n=11 post, Treatment n=13 pre and n=11 post)



Category 2.0 - Information and Analysis

Comparison of Pre- and Post-Treatment Scores on Item 2.1

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)

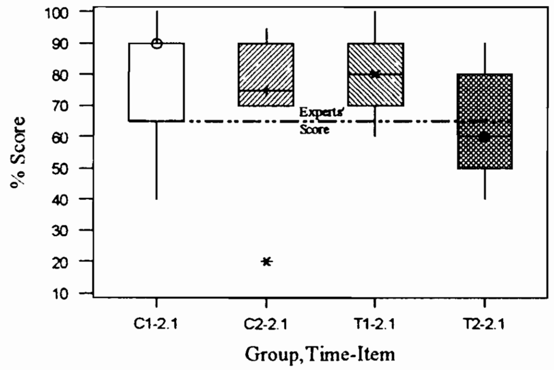
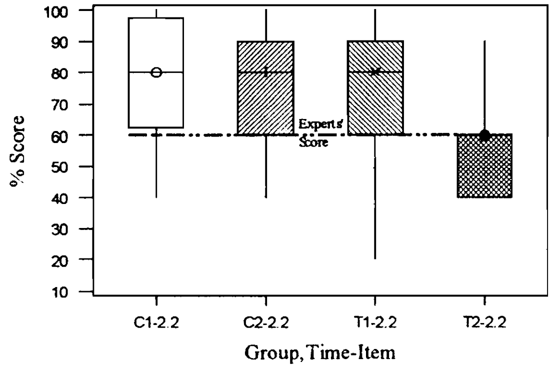


figure continues

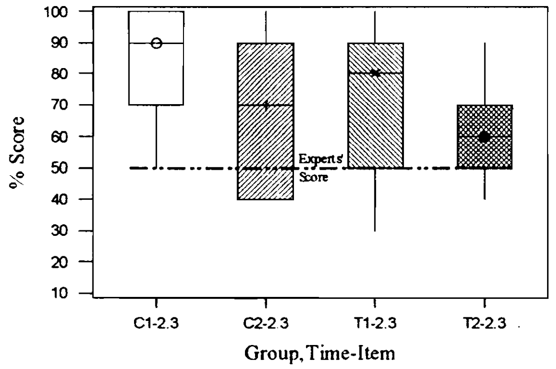
Comparison of Pre- and Post-Treatment Scores on Item 2.2

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 2.3

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 2.0

(Control n=12 pre and n=7 post, Treatment n=11 pre and n=9 post)

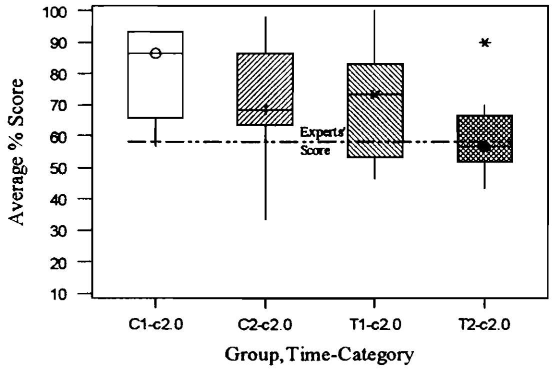
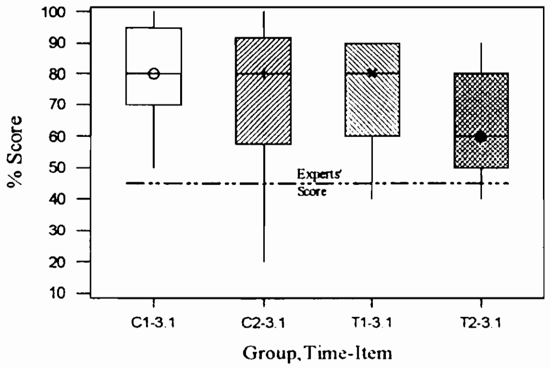


figure continues

Category 3.0 - Strategic Planning

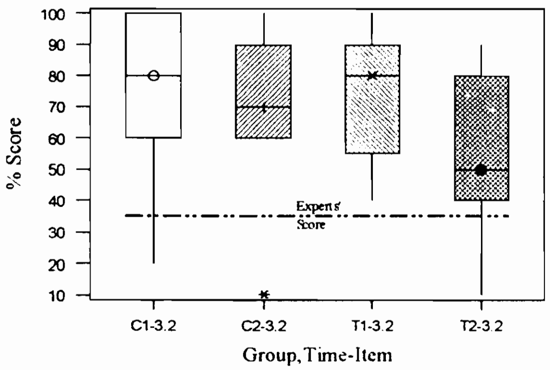
Comparison of Pre- and Post-Treatment Scores on Item 3.1

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 3.2

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 3.0

(Control n=11 pre and n=10 post, Treatment n=13 pre and n=9 post)

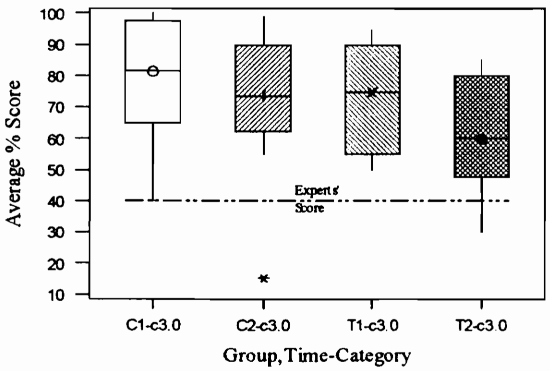
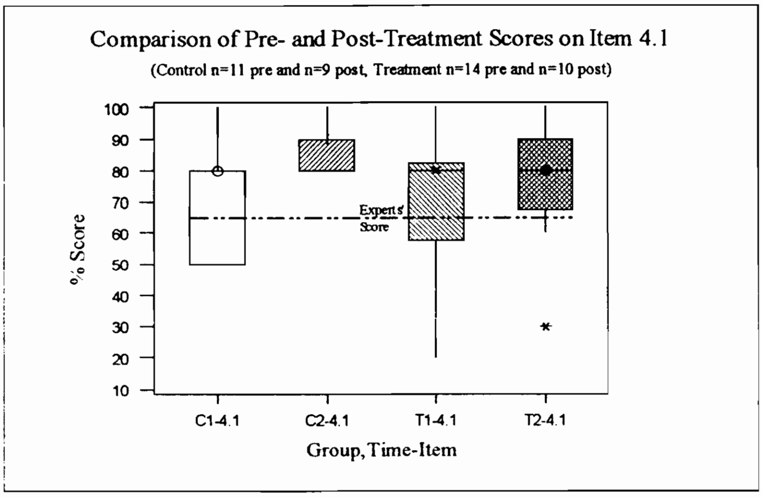
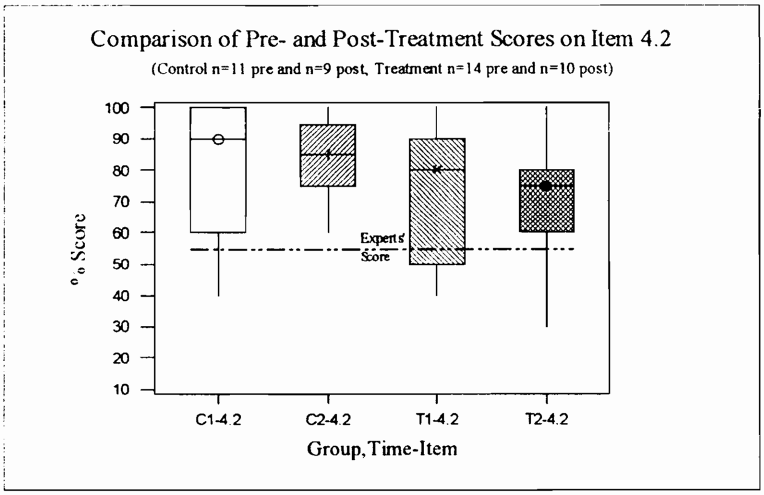


figure continues

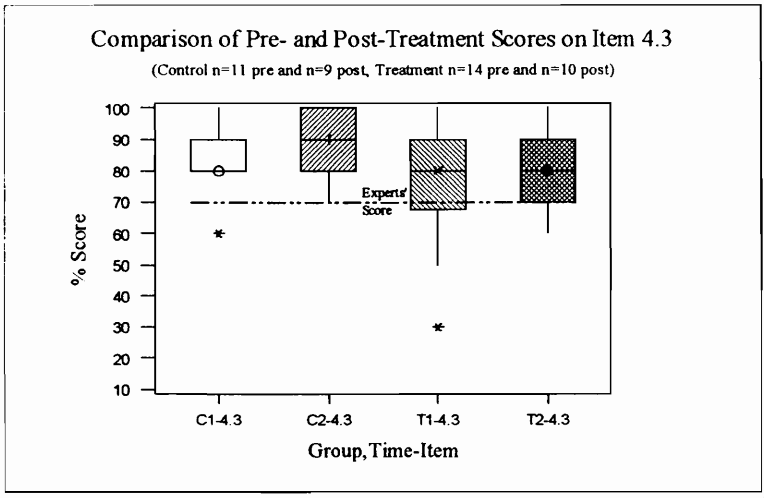
Category 4.0 - Human Resource Development and Management


Comparison of Pre- and Post-Treatment Scores on Item 4.1

(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)

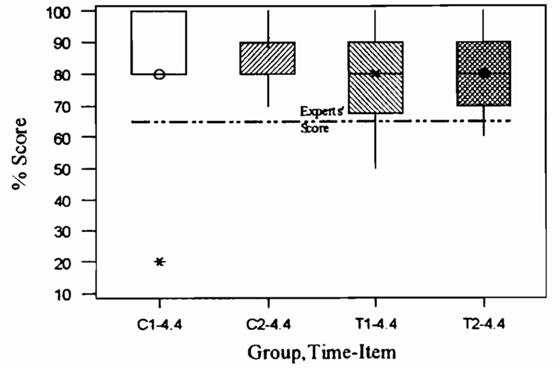

Comparison of Pre- and Post-Treatment Scores on Item 4.2

(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)


Comparison of Pre- and Post-Treatment Scores on Item 4.3

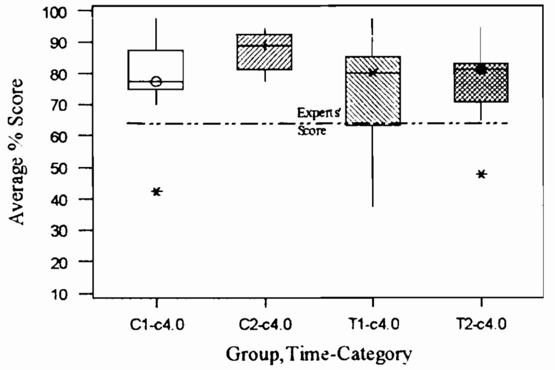
(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)

figure continues



Comparison of Pre- and Post-Treatment Scores on Item 4.4

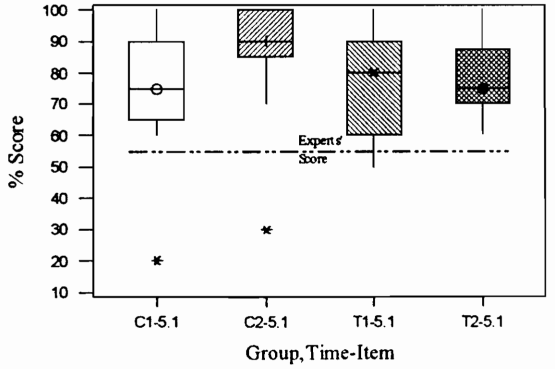
(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)



Comparison of Pre- and Post-Treatment Scores on Category 4.0

(Control n=11 pre and n=9 post, Treatment n=14 pre and n=10 post)

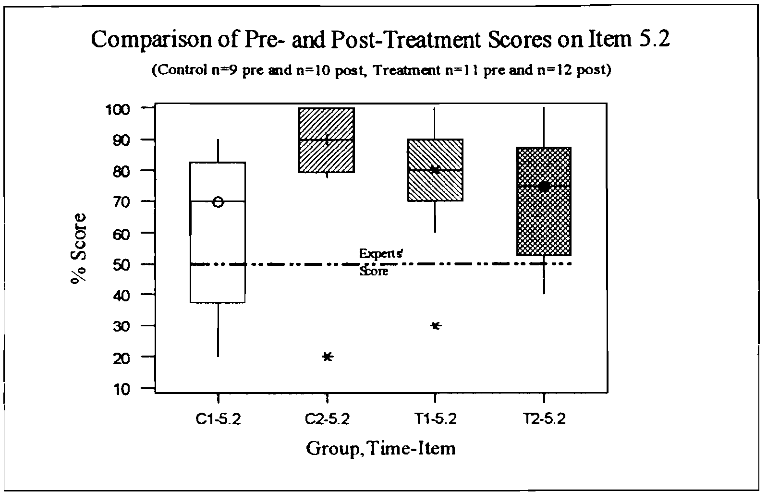
Category 5.0 - Process Management



Comparison of Pre- and Post-Treatment Scores on Item 5.1

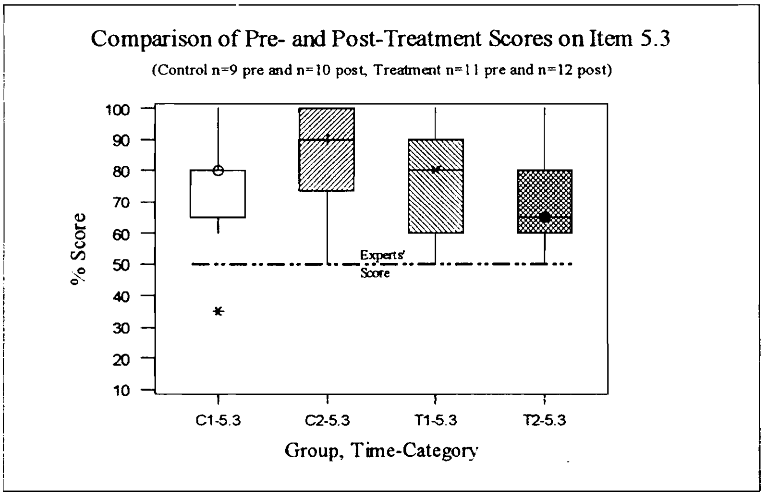
(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)

figure continues



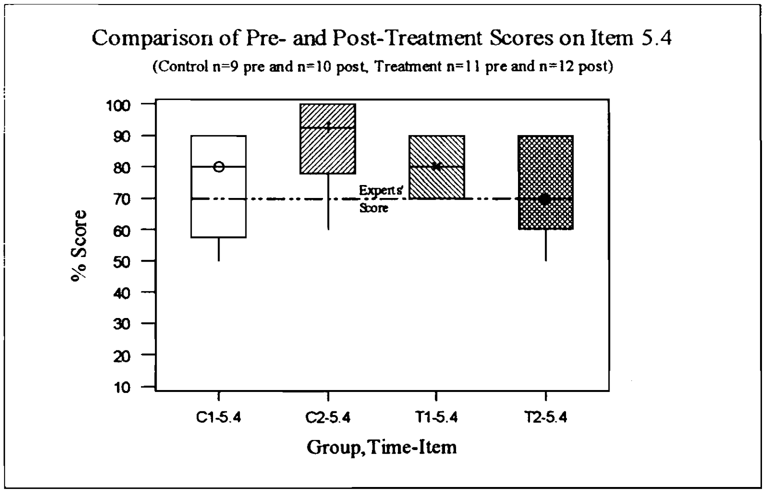
Comparison of Pre- and Post-Treatment Scores on Item 5.2

(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)



Comparison of Pre- and Post-Treatment Scores on Item 5.3

(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)



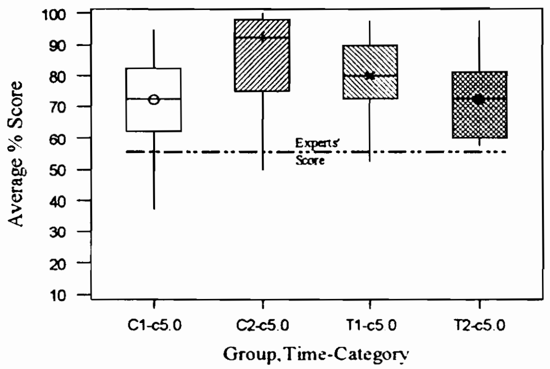
Comparison of Pre- and Post-Treatment Scores on Item 5.4

(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)

figure continues

Comparison of Pre- and Post-Treatment Scores on Category 5.0

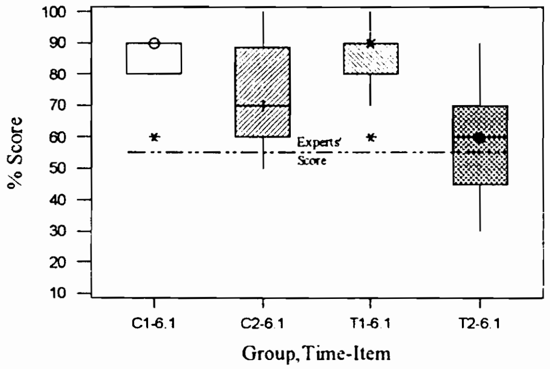
(Control n=9 pre and n=10 post, Treatment n=11 pre and n=12 post)



Category 6.0 - Business Results

Comparison of Pre- and Post-Treatment Scores on Item 6.1

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Item 6.2

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)

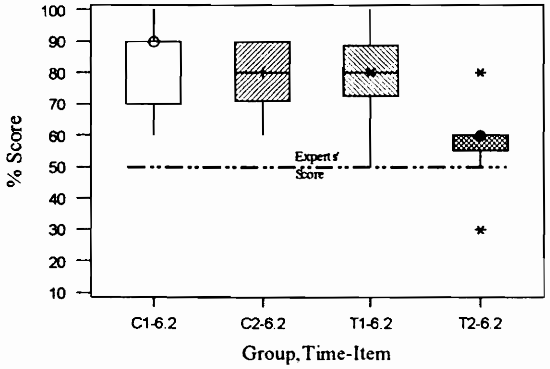
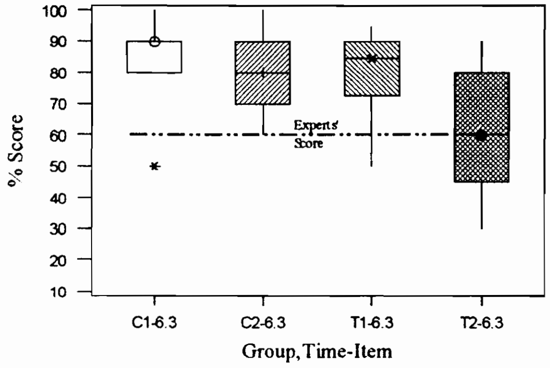


figure continues

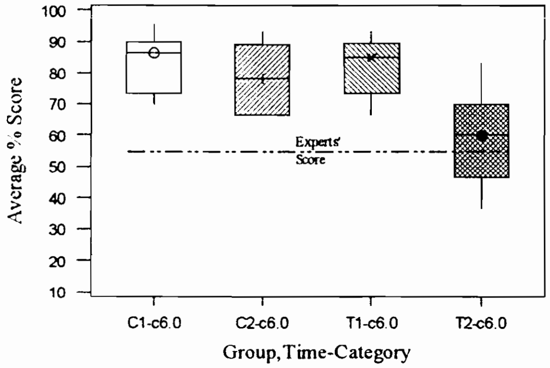
Comparison of Pre- and Post-Treatment Scores on Item 6.3

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 6.0

(Control n=7 pre and n=12 post, Treatment n=12 pre and n=9 post)



Category 7.0 - Customer Focus and Satisfaction

Comparison of Pre- and Post-Treatment Scores on Item 7.1

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)

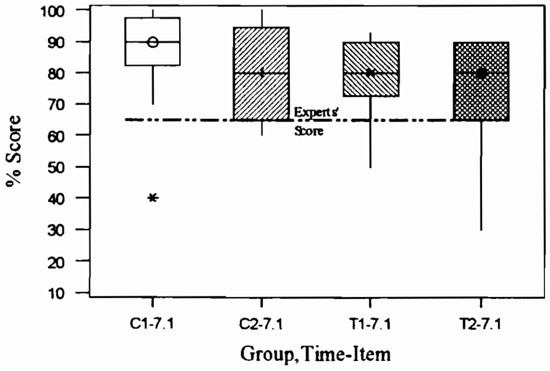
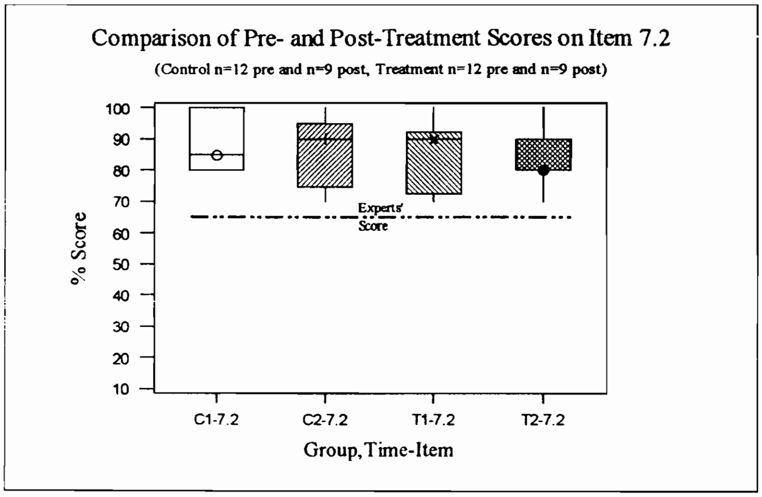
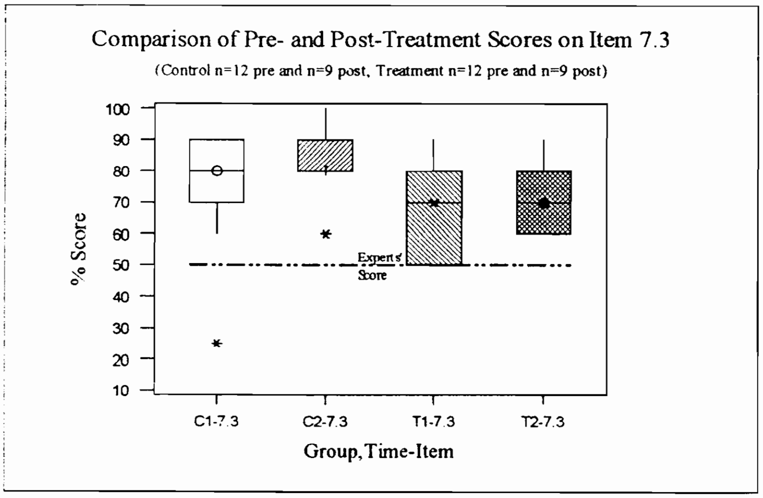


figure continues





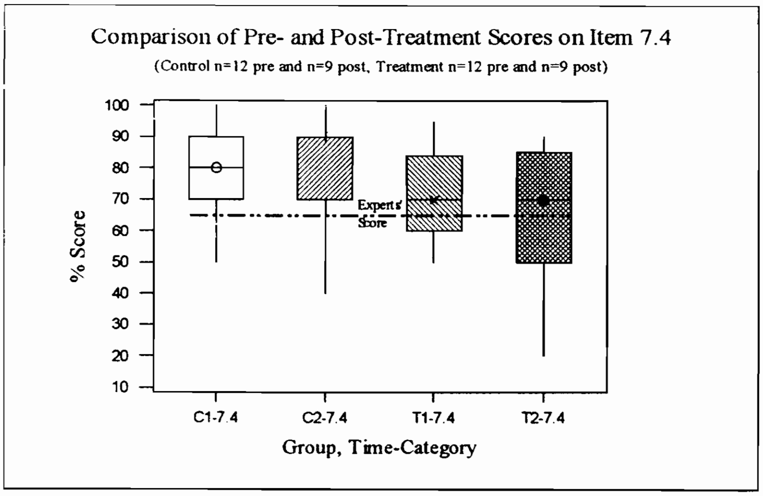
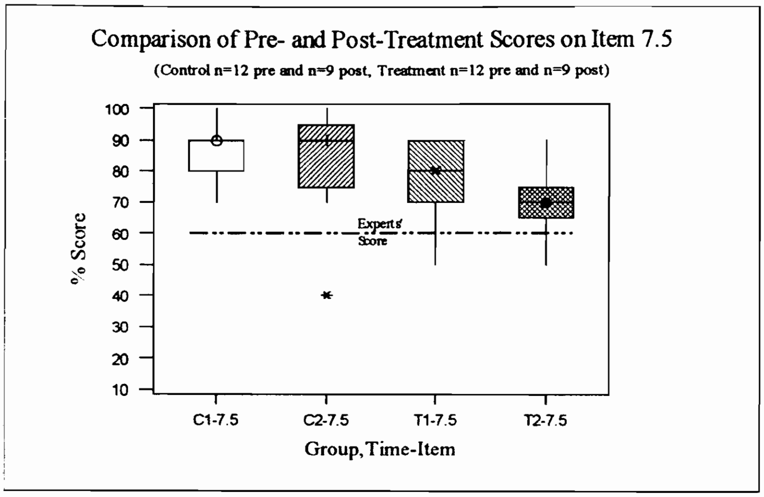
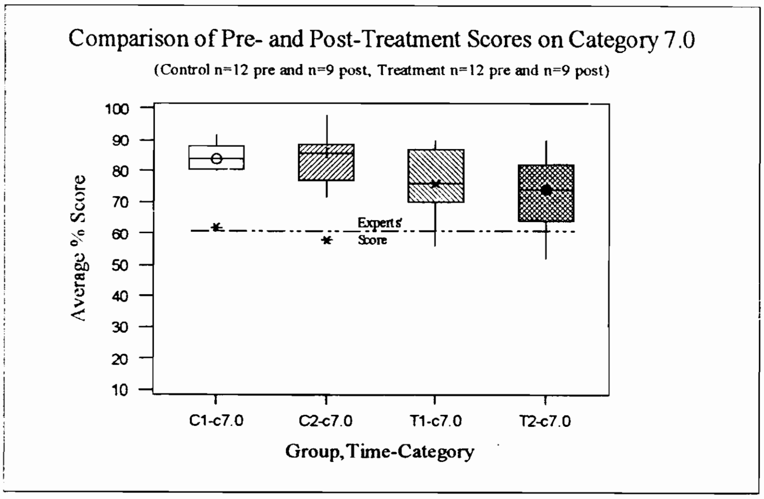


figure continues



Comparison of Pre- and Post-Treatment Scores on Item 7.5

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)



Comparison of Pre- and Post-Treatment Scores on Category 7.0

(Control n=12 pre and n=9 post, Treatment n=12 pre and n=9 post)

Content Analysis of Second Evaluation Qualitative Comments

A content analysis of the evaluators' comments on strengths, areas for improvement, and site visit issues was used to compare the experts' comments to each evaluator's comments. These comments accompanied the score of each item and were intended to provide justification for the score. The procedure used to enter and then analyze these qualitative comments was identical to that used for the first evaluation qualitative comments (see the Q5 content analysis for the procedure and examples).

The count data for hits, almost hits, and misses are contained in Appendix AU. Means were calculated for the hits, almost hits, and misses for strengths, areas for improvement, and site visit issues for each item. These means were calculated for each group and for the combined sample. Summary statistics, %Hits (%H) and %Experts (%Exp), were calculated for strengths and for all comments (i.e., strengths, areas for improvement, and site visit issues). These summary statistics were described under the Q5 content analysis.

Table 25 compares the %Hits and the %Experts of the baseline (combined groups' first evaluation), the untrained evaluators' second evaluation, and the trained evaluators' second evaluation. The second evaluation mean counts and statistics for each group by category are given in Appendix AU. The same data for the first evaluation were given in Appendix AS.

Table 25.

Comparison of %Hits and %Experts Before and After Training

Item	Group	Strengths - %Hits	Strengths - %Experts	Overall - %Hits	Overall - %Experts
1.1	Combined 1st	58%	48%	47%	30%
	Untrained 2nd	36%	27%	30%	15%
	Trained 2nd	54%	30%	43%	19%
1.2	Combined 1st	71%	56%	57%	26%
	Untrained 2nd	42%	36%	34%	15%
	Trained 2nd	68%	46%	48%	19%
1.3	Combined 1st	73%	59%	58%	25%
	Untrained 2nd	41%	37%	31%	14%
	Trained 2nd	47%	51%	36%	20%
2.1	Combined 1st	34%	48%	30%	25%
	Untrained 2nd	20%	29%	22%	18%
	Trained 2nd	26%	32%	22%	16%
2.2	Combined 1st	30%	48%	26%	22%
	Untrained 2nd	22%	31%	22%	18%
	Trained 2nd	13%	25%	10%	13%
2.3	Combined 1st	26%	38%	31%	20%
	Untrained 2nd	18%	18%	21%	15%
	Trained 2nd	28%	26%	18%	17%
3.1	Combined 1st	36%	38%	30%	15%
	Untrained 2nd	29%	35%	16%	13%
	Trained 2nd	43%	40%	34%	15%
3.2	Combined 1st	28%	28%	31%	19%
	Untrained 2nd	27%	26%	29%	19%
	Trained 2nd	17%	21%	17%	15%

table continues

Item	Group	Strengths - %Hits	Strengths - %Experts	Overall - %Hits	Overall - %Experts
4.1	Combined 1st	10%	23%	11%	10%
	Untrained 2nd	11%	20%	11%	6%
	Trained 2nd	13%	42%	12%	15%
4.2	Combined 1st	31%	30%	23%	15%
	Untrained 2nd	28%	26%	25%	11%
	Trained 2nd	17%	31%	21%	16%
4.3	Combined 1st	25%	28%	19%	14%
	Untrained 2nd	28%	27%	24%	10%
	Trained 2nd	22%	32%	19%	14%
4.4	Combined 1st	42%	44%	35%	20%
	Untrained 2nd	26%	32%	30%	14%
	Trained 2nd	31%	39%	23%	16%
5.1	Combined 1st	18%	36%	15%	15%
	Untrained 2nd	16%	38%	14%	14%
	Trained 2nd	9%	32%	9%	14%
5.2	Combined 1st	7%	15%	6%	9%
	Untrained 2nd	12%	19%	16%	10%
	Trained 2nd	3%	13%	7%	9%
5.3	Combined 1st	10%	25%	12%	15%
	Untrained 2nd	6%	14%	7%	7%
	Trained 2nd	6%	13%	8%	12%
5.4	Combined 1st	43%	40%	28%	18%
	Untrained 2nd	21%	32%	19%	15%
	Trained 2nd	28%	31%	19%	14%

table continues

Item	Group	Strengths - %Hits	Strengths - %Experts	Overall - %Hits	Overall - %Experts
6.1	Combined 1st	77%	41%	47%	16%
	Untrained 2nd	28%	27%	27%	13%
	Trained 2nd	42%	25%	46%	20%
6.2	Combined 1st	51%	43%	27%	16%
	Untrained 2nd	28%	35%	27%	16%
	Trained 2nd	30%	31%	24%	18%
6.3	Combined 1st	31%	27%	22%	13%
	Untrained 2nd	32%	30%	34%	15%
	Trained 2nd	38%	20%	41%	17%
7.1	Combined 1st	27%	35%	22%	18%
	Untrained 2nd	26%	36%	23%	16%
	Trained 2nd	28%	33%	21%	15%
7.2	Combined 1st	23%	40%	20%	16%
	Untrained 2nd	19%	29%	13%	12%
	Trained 2nd	20%	36%	20%	17%
7.3	Combined 1st	33%	32%	25%	13%
	Untrained 2nd	16%	23%	12%	9%
	Trained 2nd	13%	23%	19%	14%
7.4	Combined 1st	32%	25%	23%	13%
	Untrained 2nd	33%	24%	22%	13%
	Trained 2nd	42%	26%	23%	13%
7.5	Combined 1st	22%	14%	18%	6%
	Untrained 2nd	52%	26%	42%	9%
	Trained 2nd	44%	24%	21%	9%

Question 11 - How might the training of evaluators be improved?

Question 11 (Q11) was intended to capture the lessons learned from the training intervention portion of the experiment and from answering the previous research questions. Question 11 was also addressed by testing hypotheses 11 (H11) and 12 (H12). Following completion of the second evaluation, subjects were asked to rate each category they evaluated based on their perception of the difficulty of evaluating the category. A four point scale was used to rate perceived difficulty: easy, somewhat easy, somewhat difficult, and difficult. Subjects were also asked to rate each category they evaluated based on how close to the experts' scores they felt their scores were (i.e., perceived accuracy). A four point scale was used to rate closeness to expert scores: remote, somewhat remote, somewhat close, and close. Data used for testing H11 and H12 also provided additional evidence for addressing Q11. Categories perceived as more difficult to evaluate may indicate a need for improved or increased training. Categories with larger gaps between perceived and actual accuracy may also indicate a need for improved or increased training.

H11: Evaluators' perceived difficulty of evaluating a category will be negatively related to their accuracy in scoring that category.

Hypothesis 11 tested for a relationship between the evaluators' perceived difficulty of evaluating a category and the evaluators' accuracy in scoring that category. It was believed that evaluators who perceived a category to be more difficult to evaluate would be less accurate in scoring that category. Thus, a negative relationship was expected for H11.

H12: Evaluators' perceived accuracy in scoring a category will be positively related to their accuracy in evaluating that category.

Hypothesis 12 tested for a relationship between the evaluators' perceived accuracy in scoring a category and the evaluators' actual accuracy in scoring that category. It was believed that evaluators who felt their scores were closer to the experts would be more accurate in scoring that category. Thus, a positive relationship was expected.

Hypotheses 11 and 12 were tested using the accuracy statistics from Q5 and Q10, matched with the subjects' responses to the questions regarding perceived difficulty and perceived accuracy. Since there were no differences in the accuracy of pre-treatment (i.e., first evaluation) scores, the control and treatment groups' first evaluation scores were combined to obtain increased sample size. For comparison, H11 and H12 were also tested using the accuracy statistics from the treatment group's second evaluation scores; however, the small sample size limited the usefulness of these data.

Testing Hypothesis 11

An ordered categories procedure was used to test H11 for both the combined first evaluation scores and the treatment group's second evaluation scores. The ordered categories procedure is a form of correlation analysis where both variables are classified into discrete categories with a natural order. The test can be used to determine if a positive, negative, or no relation appears between the variables. For each of the seven categories, separate tests were conducted using elevation and dimensional accuracy as indices of accuracy compared to the subjects' rating of perceived difficulty in evaluating that category.

Perceived difficulty was divided into four categories corresponding to the response choices: easy, somewhat easy, somewhat difficult, and difficult. Actual accuracy was divided into two categories, better and worse. A four by two matrix was created by placing perceived difficulty in the rows and actual accuracy in the columns. A data reduction procedure (discussed next) was used to determine the number of observations in each cell of the ordered categories matrix.

A spreadsheet was constructed for each evaluation category (e.g., 1.0, 2.0) and group (combined groups' first evaluation or treatment group's second evaluation). The spreadsheet contained the accuracy statistics for all the subjects who evaluated that category. Each row of the spreadsheet contained the subject number, dimensional accuracy index, elevation index, and perceived difficulty[22] for one subject. The rows were sorted in ascending order by dimensional accuracy indices. Subjects were then categorized as either having better or worse accuracy compared to all the subjects who evaluated that category[23]. Those with a DA index below the median (a lower index represents better accuracy) were identified as having better accuracy. Those with a DA index above the median were identified as having worse accuracy. If a subject's DA index was equal to the median, then it was compared to the mean. If the DA index was greater than the mean, then that subject was identified as having worse accuracy. If the DA index was equal to the median and less than the mean, then that subject was identified as having better accuracy. The number of subjects rating the perceived difficulty at each level (easy, somewhat easy, somewhat difficult, or difficult) were then counted for each level of accuracy (better or worse). The results of this count were entered into a 4 x 2 ordered categories matrix to be tested for the hypothesized relationship. Next, the rows were sorted by elevation indices and the same procedure followed. Data summary worksheets were used to summarize the results of this data reduction procedure for each evaluation category and group.

Tested Hypotheses

The ordered categories procedure tested the following hypothesis for each category, group, and accuracy index.

H_0: Evaluators' perceived difficulty of evaluating category X.0 and their actual accuracy in evaluating category X.0 are independent.

H_1: Evaluators' perceived difficulty of evaluating category X.0 and their actual accuracy in evaluating category X.0 are negatively related.

Where X.0 may be any category from 1.0 to 7.0 and actual accuracy is measured by elevation or dimensional accuracy. Appendix AM contains an example SAS output from

[22] Perceived difficulty values were taken from the questionnaire responses provided by the subjects.

[23] All the subjects' accuracy indices were used for the purpose of ranking the subjects based on accuracy, even though a few of these subjects did not provide usable responses to the question regarding perceived difficulty. All the subjects refers to those used for testing H5 and H10.

the H11 ordered categories tests. Appendix AN describes the test statistic and rejection region used for testing H11.

Test Results

The ordered categories test results for H11 are summarized in Table 26. The table includes the mean perceived difficulty in evaluating each category, providing some indication of which categories the subjects found most difficult to evaluate.

Table 26.

Summary of the Ordered Categories Test Results for Hypothesis 11

Group[24]-Category	Mean Perceived Difficulty	Test Statistic for Perceived Difficulty and Dimensional Accuracy	Significant?	Test Statistic for Perceived Difficulty and Elevation	Significant?
CT-1.0	2.50	-0.346	no	1.455	no
CT-2.0	2.95	0.098	no	2.222	no
CT-3.0	2.88	-0.513	no	1.997	no
CT-4.0	3.15	0.337	no	-0.337	no
CT-5.0	3.50	-0.944	no	2.478	no
CT-6.0	2.60	0.128	no	-1.713	yes
CT-7.0	2.57	0.708	no	-0.159	no
T2-1.0	2.78	0.588	no	0.258	no
T2-2.0	3.13	-3.325	yes	-0.547	no
T2-3.0	2.75	-1.442	no	-0.354	no
T2-4.0	3.14	-0.747	no	-0.747	no
T2-5.0	2.82	-1.559	no	2.475	no
T2-6.0	2.44	0.000	no	1.282	no
T2-7.0	2.63	0.759	no	0.340	no

[24] Groups for this test were the combined control and treatment groups' first evaluation (CT) and the treatment group's second evaluation (T2). Usable responses per category ranged from 15 to 23 for CT and from 7 to 11 for T2.

Testing Hypothesis 12

An ordered categories procedure was used to test H12 for both the combined first evaluation scores and the treatment group's second evaluation scores. The ordered categories procedure was identical to that used to test H11, except that perceived accuracy rather than perceived difficulty was compared to actual accuracy.

Perceived accuracy was divided into four categories corresponding to the response choices: remote, somewhat remote, somewhat close, and close. Actual accuracy was divided into two categories, better and worse. A four by two matrix was created by placing perceived accuracy in the rows and actual accuracy in the columns. A data reduction procedure, like the one used for H11, was used to determine the number of observations in each cell of the ordered categories matrix.

Tested Hypotheses

The ordered categories procedure tested the following hypothesis for each category, group, and accuracy index.

H_0: Evaluators' perceived accuracy in evaluating category X.0 and their actual accuracy in evaluating category X.0 are independent.
H_1: Evaluators' perceived accuracy in evaluating category X.0 and their actual accuracy in evaluating category X.0 are positively related.

Where X.0 may be any category from 1.0 to 7.0, perceived accuracy is how close the subject felt their scores were to the experts' scores, and actual accuracy is measured by elevation or dimensional accuracy. Appendix AN describes the test statistic and rejection region used for testing H12.

Test Results

The ordered categories test results for H12 are summarized in Table 27. The table includes the mean perceived accuracy in evaluating each category, providing some indication of which categories the subjects felt they evaluated more accurately. The small sample sizes of the treatment group's second evaluation occasionally produced multiple cells with zero observations, thus resulting in an undefined value for the test statistic.

Table 27.

Summary of the Ordered Categories Test Results for Hypothesis 12

Group[25]-Category	Mean Perceived Accuracy	Test Statistic for Perceived Accuracy and Dimensional Accuracy	Significant?	Test Statistic for Perceived Accuracy and Elevation	Significant?
CT-1.0	2.42	0.917	no	-1.993	no
CT-2.0	2.32	-2.933	no	1.362	no
CT-3.0	2.40	1.741	yes	-0.514	no
CT-4.0	2.58	1.277	no	0.750	no
CT-5.0	2.33	0.320	no	-2.000	no
CT-6.0	2.93	0.478	no	-1.736	no
CT-7.0	2.65	-0.492	no	1.292	no
T2-1.0	2.78	-1.732	no	0.000	no
T2-2.0	2.86	undefined	no	undefined	no
T2-3.0	2.43	-0.474	no	undefined	no
T2-4.0	2.57	undefined	no	-1.247	no
T2-5.0	2.55	-0.462	no	undefined	no
T2-6.0	3.00	0.000	no	undefined	no
T2-7.0	2.71	-2.193	no	2.193	yes

[25] Groups for this test were the combined control and treatment groups' first evaluation (CT) and the treatment group's second evaluation (T2). Usable responses per category ranged from 14 to 20 for CT and from 7 to 11 for T2.

Question 12 - Which evaluator characteristics best predict the accuracy of the evaluators' scores?

Question 12 refers to the evaluator characteristics identified by questionnaire following the second evaluation. The questionnaire was shown in Appendix N. The characteristics measured by the questionnaire are summarized in Table 28. No hypotheses were tested for question 12. Exploratory regression analyses were used to identify which of these characteristics were the best predictors of the accuracy of the evaluators' scores. Both elevation and dimensional accuracy were used to see if the predictors differed for these two indices. Since the purpose was to look for relationships between evaluator characteristics and accuracy, first evaluation (pre-training) accuracy indices were used for the dependent variable. Mean and median responses were calculated to assist with data reduction and to provide additional information about the subjects.

Table 28.

Evaluator Characteristics Measured by the Questionnaire.

Characteristics	Scale
Prior use of MBNQA criteria	yes/no
How the MBNQA criteria were used	15 possible uses listed
Participation in other forms of organizational assessments	open ended (w/suggested categories)
Highest degree obtained	<Bachelor(1), Bachelor(2), Master(3)
Educational specialty of highest degree	6 categories or other
Days of Quality & Productivity Improvement Training in past 10 years	number of days (classroom, on the job, and conference sessions)
Years of professional work experience	number of years
Current Job Function	11 categories including student and unemployed
Years of work experience in a quality control, quality assurance, or quality improvement function	number of years
Supervisory Responsibility	yes/no
Number or employees supervised	number of employees
Description of current employer's industry	6 categories
Total number of employees in your employing organization	number of employees (later categorized, govt. included in largest category)
Your Current Age	years
Gender	male/female

The data from the questionnaire were entered into a spreadsheet for preliminary analysis and data reduction. Sixty-seven usable questionnaires were returned by those subjects whose score data passed the screening conducted prior to calculation of the accuracy indices (see Question 5). This represented a usable response rate of 82.7% (67 out of 81). The mean and median were calculated for each characteristic measured on an ordinal, interval, or ratio scale. Characteristics measured on a nominal scale were entered as indicator or dummy variables (i.e., a subject either was or was not in that category, represented by a one or zero). This allowed the calculation of the percent of the subjects falling into each category. Once in this form, the data were reviewed to identify underutilized categories or continuous variables with dispersions that might be better suited for an ordinal scale. The following paragraphs describe the results of this review.

Evaluator Characteristics

Only one subject (1.5%) had used the MBNQA criteria prior to this study. Given this small percentage, how they had used the MBNQA criteria was deemed irrelevant. A different subject (again, 1.5%) had been involved in applying for a quality or productivity award other than the Baldrige. Eight subjects (11.9%) had been involved with ISO 9000 certification. Ten subjects (14.9%) had been involved in some other form of organizational assessment, including Military Standards, supplier certification, and self-assessment. Overlap between those involved with quality and productivity award criteria, ISO 9000, or other organizational assessments resulted in a total of 20.9% of the subjects having had some prior experience with organizational assessment criteria. These categories were all combined into a single category regarding prior experience with organizational assessment criteria (variable name = assess). Those with prior experience were assigned a value of one, those without prior experience were assigned a value of zero.

Over 19% of the subjects had completed a master's degree. Only one subject had less than a bachelor's degree. The remainder of the subjects (79%) had completed a bachelor's degree. Several of these noted that they were within one semester of completing a master's degree. The variable name "degree" was used to indicate the highest degree completed.

The majority (79%) of the subjects had completed their highest degree in engineering. No other specialty comprised more than 6% of the subjects. Both business and physical sciences were identified as the highest degree completed by 6% of the subjects. 3% of the subjects had completed their highest degree in math or statistics and 3% had completed their highest degree in a specialty not listed (e.g., computer science). One subject had completed their highest degree in a social science and one subject in education. Due to the dominance of a single specialty, these categories were combined into a single category representing whether the highest degree completed was in engineering (variable name = engrdeg). Those whose highest degree completed was in engineering were assigned a value of one and all others were assigned a value of zero.

The days of quality and productivity improvement training in the past ten years were combined into a single value (variable name = qptrng). That is, days of classroom training (other than formal education), on-the-job training, and conference session attendance were combined. Eighteen subjects (26.9%) indicated they had not received any quality and productivity training in the past ten years. The median amount of quality and productivity training was eight days. The mean was 48 days, skewed by two outliers. One subject with over six years work experience in the QA/QC function claimed nearly all this time as on-the-job training. This subject had experience with organizational assessments and provided no other questionable responses. A second subject with 20 years of professional experience indicated 280 days of quality and productivity training in the past ten years. This subject provided no reason to question this response. If the first subject's response were changed to equal the second highest response (e.g., 280 days), the mean would drop to less than 30 days. Given the potential for opening Pandora's box by making assumptions about subjects' responses, neither of these were discarded. Coding the data into ordinal categories resulted in a noticeable loss of predictive power. The values for this variable were entered as given.

Nearly all the subjects (92.5%) had one or more years of professional work experience (variable name = exp). The mean work experience was 7.8 years and the median was 5.5 years. One subject reported 26 years of work experience and six others reported 20 or more years (i.e., 10.4% with 20 or more years of work experience). Most of the subjects (71.6%) did not have experience working in the quality control/quality assurance/quality improvement (QC/QA/QI) function (variable name = yrsqcqa). The mean experience in the QC/QA/QI function was 1.1 years and the median was zero. One subject had 15 years experience in QC/QA/QI and the second most experience in QC/QA/QI was 8 years. The values for both of these variables were entered as given.

Subjects were asked to identify their current job function and given eleven categories as examples. Nearly all subjects chose one of these eleven categories. Those that did not choose one of the given categories typically provided a more specific job title that was easily categorized. No subjects identified their current job function as human resources, finance/accounting, sales/marketing, or teaching/training. These categories were discarded. Nearly 12% of the subjects identified their current job function as executive/administrative. Only 3% identified their job function as production/service delivery and only 1.5% as maintenance. Due to these small percentages, production/service delivery and maintenance were combined with the 7.5% who chose quality control/quality assurance into a category called "operations." Only 7.5% identified research and development as their job function. These were combined with the 37.3% in engineering to produce a category called "technical." Twenty-eight percent of the subjects identified their function as full-time student; however, a number implied that they were employed in teaching or research capacities. This resulted in four current job function categories: executive/administration (exec); operations (oper), technical (tech), and full-time student (std). Each subject was coded as either being in a category (i.e., assigned a value of one) or not (i.e., assigned a category of zero). No subject was in more than one category. A single subject indicated they were unemployed and they were coded with all zeros for the job function categories.

Nearly 27% of subjects indicated they currently have supervisory responsibility. The number of employees supervised ranged from 2 to 220. The mean number of employees supervised among all 67 subjects was 6.4 employees, the median was zero. The mean number of employees supervised by those with supervisory responsibility was 23.8 employees and the median was 6 employees. Having supervisory responsibility and the number of employees supervised were combined by simply entering the number of employees supervised (variable name = supv). Those without supervisory responsibility were coded as supervising zero employees. Coding the data into ordinal categories resulted in a noticeable loss of explanatory power. The values for this variable were entered as given.

Subjects were asked to identify which of 6 categories best described their current employer: manufacturing; service; federal government; state or local government; education; and health care. Only one subject described his/her employer as health care. This category was combined with service, resulting in 22.4% of the subjects being identified as employed in the service industry (variable name = svc). Nearly 15% of the subjects described their employer as manufacturing (variable name = mfg). The most common employer was the federal government (variable name = fed), with 31.3% of the subjects. A number of subjects identified their employer as education. Some of these were employed by the university, while others were employed in secondary education (e.g., by local school boards). Several of those who identified their employer as education, identified themselves as full-time students. Others who identified themselves as students identified their employer as state or local government. To reduce the probability of category overlap, education and state or local government were combined into a single category (variable name = stloc) which contained 22.4% of the subjects. Each subject was coded as either being in a category (i.e., assigned a value of one) or not (i.e., assigned a category of zero). No subject was in more than one category. Subjects that did not describe an employer (approximately 9%) were coded with all zeros for the employer's description categories. These subjects were either full-time students or unemployed.

Subjects were asked to list the total number of employees in their entire company or employing organization. Those employed by the federal or state government were asked to simply write "govt." rather than provide a number. Those that did not describe an employer (9%) were coded as working for an organization with zero employees. After reviewing the data, the following categories were developed for the variable named size. Those working for organizations with between 1 and 500 employees (14.9%) were coded with a value of one. Those working for organizations with between 501 and 5000 employees (7.5%) were coded with a value of two. Those working for organizations with between 5001 and 10000 employees (26.9%) were coded with a value of three. Those working for organizations with more than 10000 employees or federal, state, or local government (40%) were coded with a value of four.

The mean age of the subjects was 30.3 years, with a median of 29 years. Subjects' ages ranged from 22 to 49. The values for this variable were entered as given. Over twenty-two percent (22.7%) of the subjects that reported gender were female. This variable (name = gndr) was entered by coding females with a value of one and males with a value of zero.

Exploratory Data Analysis

Exploration of the data began with the examination, consolidation, and calculations of central tendency described above. The data were entered into Minitab for further analysis. Descriptive statistics were calculated for each variable, including the dependent variables (see Appendix AO). A correlation analysis was performed to see which variables were most strongly related. Table 29 contains the correlation table. Appendix AY contains observations following a detailed review of the correlation table.

Table 29.

Correlations of Independent and Dependent Variables used in Q12 Regression Analyses

	Eavg	DAavg	qptrng	exp	yrsqcqa	supv	age	degree
DAavg	-0.435							
qptrng	-0.207	-0.021						
exp	0.233	-0.095	0.012					
yrsqcqa	-0.014	0.166	0.301	0.272				
supv	-0.163	0.161	-0.040	0.455	-0.057			
age	0.229	-0.186	0.020	0.958	0.218	0.429		
degree	-0.003	0.149	0.002	0.068	0.018	0.282	0.106	
size	0.081	0.045	0.135	0.380	0.203	0.057	0.315	0.115
engreduc	-0.212	-0.002	0.078	-0.302	0.040	-0.269	-0.267	-0.218
assess	0.072	0.074	0.271	0.162	0.219	0.032	0.123	0.130
gndr	0.023	0.115	-0.105	-0.115	0.007	-0.069	-0.196	-0.087
exec	0.138	0.008	-0.079	0.476	-0.041	0.462	0.486	0.172
oper	-0.135	0.364	0.321	0.118	0.546	-0.015	-0.009	-0.157
tech	0.026	-0.074	-0.110	0.038	-0.146	-0.149	0.047	-0.169
std	-0.022	-0.125	-0.057	-0.409	-0.173	-0.143	-0.364	0.283
mfg	0.066	0.033	0.333	0.279	0.196	-0.000	0.256	-0.079
svc	0.010	-0.078	-0.078	-0.023	-0.004	0.160	-0.058	0.027
fed	0.122	0.023	-0.097	0.297	0.069	0.002	0.277	-0.058
stloc	-0.171	0.135	-0.071	-0.339	-0.146	-0.113	-0.295	0.282

	size	engreduc	assess	gndr	exec	oper	tech	std
engreduc	-0.191							
assess	0.075	-0.007						
gndr	-0.020	-0.072	-0.105					
exec	0.019	-0.264	0.037	0.131				
oper	0.261	-0.037	0.037	0.131	-0.136			
tech	0.312	0.020	0.054	0.013	-0.332	-0.332		
std	-0.399	0.160	-0.079	-0.185	-0.232	-0.232	-0.567	
mfg	0.011	0.215	0.300	-0.128	-0.025	0.104	0.212	-0.264
svc	-0.234	0.012	-0.012	0.224	0.133	-0.087	0.092	-0.100
fed	0.659	-0.207	-0.110	-0.060	0.049	0.148	0.233	-0.354
stloc	-0.043	-0.076	-0.012	-0.122	-0.087	-0.087	-0.340	0.536

	mfg	svc	fed
svc	-0.225		
fed	-0.283	-0.363	
stloc	-0.225	-0.288	-0.363

Each variable was used as the independent variable in a simple linear regression analysis (SLR) where accuracy was the dependent variable. This was done using elevation as the measure of accuracy and repeated using dimensional accuracy as the measure of accuracy. This provided some indication of which characteristics had predictive power over accuracy. Fits were plotted, unusual observations were dropped, and residuals were examined to see if this variable required further refinement before entering the subsequent multiple regression analyses. The results of this procedure are summarized in Tables 30 and 31. When elevation was used as the dependent variable, observation 26 was consistently identified as an outlier regardless of the independent variable. This observation was dropped for the multiple regression analysis when elevation was the dependent variable. When dimensional accuracy was the dependent variable, observation 29 was consistently identified as an outlier, in many cases an extreme outlier. Observations 1, 5, and 30 were frequently identified as outliers, too. Observation 29 was dropped for the multiple regression analysis when dimensional accuracy was the dependent variable.

Table 30.

Potential Variables for Stepwise Regression when Elevation is the Dependent Variable

Original Variable	Recommended Predictors	Comments from SLR
qptrng	drop or use	includes extreme outlier (1514 days)
	qptrng5E	drops extreme outlier and all zero values very little prediction; Rsq = 0.007
exp	use or replace with	large influence for values 25 or greater Rsq = 0.054
	expE1	four data points 25 or greater dropped Rsq = 0.082
yrsqcqa	drop or use	lots of zero values
	YrsQC3	drops two large influences (yrs = 15 & 8) and all zero values; no unusual obs., Rsq = 0.080
supv	use or drop?	lots of zero values & a few extreme outliers Rsq = 0.027
age	use or drop due to multicollinearity	oldest subjects have large influence, dropping made little difference; Rsq = 0.052
degree	drop or use	little or no prediction
	degree2	drops the only obs. with less than a bachelor's
size	drop or use	little prediction with zero size included
	size4E	drops zero and one (small) size Rsq = 0.065
engrdeg	use	Rsq = 0.045
assess	drop?	little prediction, random looking dispersion
gndr	drop?	little prediction, random looking dispersion
exec	use or drop	exec obs. have large influence Rsq = 0.019
oper	use or	oper obs. have large influence; Rsq = 0.018
	oper3E	drops the only obs. with large st. residual and large influence; Rsq = 0.055
tech	drop	little prediction, random looking dispersion
std	drop	little prediction, random looking dispersion
mfg	drop	little prediction, dropping obs. with large std. residuals and large influence reduces mfg. obs. sample size by 2 out of 9 and appears arbitrary (points are similar)
svc	drop	no prediction, random looking dispersion
fed	use or drop?	Rsq = 0.015
stloc	use or drop?	Rsq = 0.029

Table 31.

Potential Variables for Stepwise Regression when Dimensional Accuracy is the Dependent Variable.

Original Variable	Recommended Predictors	Comments from SLR
qptrng	drop or use	includes extreme outlier (1514 days)
	qptrng5	drops two extreme outliers, all zero values, and three obs. w/large st. residuals; Rsq = 0.017
exp	drop or use	obs. 29 fails residual analysis tests
	exp2	drops obs. 29, large influence for values 25 or greater; Rsq = 0.023
	exp4	drops obs. 29, values 25 or greater, and four obs. with large residuals; Rsq = 0.082
yrsqcqa	drop or use	lots of zero values, but relatively tight dispersion, obs. 29 fails residual analysis tests; Rsq = 0.028
	YrsQC5	drops obs. 29, and two others (obs. 5 & 30) w/subsequent large residuals (>2.7); Rsq = 0.057
supv	drop or use	includes extreme outlier (220 employees)
	supv7D	drops middle mgrs. (obs. 29, 46,& 52), obs. 30 w/large st. residual, and zero values; possible sample size problem: Rsq = 0.538
age	drop or use	obs. 29 fails residual analysis tests
	age3D	drops obs. 29 and obs. 30 (subsequent 2.82R); Rsq = 0.051
	age5D	drops add'l unusual obs. 1, 5, 17, 43, &46; Rsq = 0.109
degree	drop	poor predictor
size	drop or use	obs. 29 fails residual analysis tests, sample size problem with size = zero
	size3D	drops obs. 29 and those with value = zero, dropping add'l. unusual obs. reduces predictive power; Rsq = 0.024
engrdeg	drop or use	obs. 29 fails residual analysis tests, obs. 1, 5, 29, & 30 form a cluster of outliers
	engrdeg2	drops obs. 1, 5, 29, & 30, weak predictor; Rsq = 0.011
assess	drop	poor predictor
exec	drop	exec obs. have large influence, poor predictor
oper	use or	oper. obs. have large influence, obs. 5 & 30 have large residuals; Rsq = 0.133
	oper 2	drops obs. 5 & 30; Rsq = 0.189
tech	drop or use	obs. 29 fails residual analysis tests, obs. 5 & 30 also outliers
	tech2	drops obs. 5, 29, & 30; Rsq = 0.022
std	drop	obs. 29 fails residual analysis tests, obs. 5 & 30 also outliers, dropping unusual obs. reduces predictive power; Rsq = 0.016
mfg	drop or use	mfg. obs. have large influence, obs. 29 fails residual analysis tests
	mfg4	drops obs. 1, 5, 29, & 30; Rsq = 0.016
svc	drop or use	obs. 5 & 29 are outliers; Rsq = 0.006
	svc3	drops obs. 5, 29, & 30; Rsq = 0.016
fed	drop	poor predictor
stloc	drop or use	obs. 29 fails residual analysis tests; Rsq = 0.018
	stloc2	drops obs. 29, obs. 5 now an outlier; Rsq = 0.029
	stloc5	drops subsequent outliers obs. 1, 5, & 30; Rsq = 0.027

Stepwise Regression Analyses

Stepwise multiple regression was used to identify subsets of independent variables that provided relatively high predictions of accuracy. The following procedure used elevation and then dimensional accuracy as the dependent variable. Minitab was used to conduct step-wise multiple regression. The first run used all the independent variables in their original form. The second run used all the independent variables, deleting those observations identified as extreme outliers. The third and subsequent runs used the variables from the second run, replacing them one at a time with the corresponding recommended or alternate variable (shown in Tables 30 and 31). In some cases, the recommendation was to drop the variable altogether. After each run, the original variable was reinstated and the next variable was replaced for the subsequent run. Dropping variables not only gave an indication of the dropped variable's contribution (or lack thereof) to a prediction equation, but also indicated the impact of missing observations for that variable. That is, the variable may never enter the prediction equation, yet the R-squared value could change due to inclusion of an observation that was missing for that variable only. Minitab ignores observations with a missing value for any of the variables being considered.

The next set of stepwise regression runs also used the variables from the second run. Like before, variables were replaced one at a time with the corresponding recommended or alternate variable. If an alternate variable produced better results (i.e., more prediction) than the original, it was retained for the next run. If not, the original variable was reinstated. This was repeated until each variable had been replaced or retained. Comment statements were used to insert descriptions of alternate variables and potential explanations of results into the Minitab session files. Observations during each run were documented (see Appendix AZ) for future reference. These documentation procedures facilitated comparisons across analyses. Combinations of original and alternate variables were chosen based on holistic reviews of the previous results. Maximizing sample size and improving generalizability were considered when choosing between prediction equations. The end result was the subset of original and alternate variables that provided the best prediction (based on this procedure, generalizable to those training to be evaluators) for the given sample.

Proposed Regression Equations

The following equation appeared to best predict average elevation accuracy for a sample representative of evaluators in training. This equation was based on data from 42 subjects and yielded an R-squared (adjusted) of 0.214 with three predictors.

Eavg = 4.33 - 6.46 engrdegE + 8.68 degree2 + 0.527 expE2

Where engrdegE represents whether or not the last degree completed was an engineering degree (if yes, value = 1). Four observations were omitted by using engrdegE rather than the original engrdeg. The first was the extreme outlier (obs. 26) and the other three were originally omitted due to missing values for other variables (obs. 19, 47, and 60). The single subject with less than a bachelor's degree was omitted by using degree2 rather than degree. For the variable degree2, a value of 2 represents a bachelor's degree and a value of 3 represents a master's degree as the highest degree completed. Interpretation of the equation can be simplified by transforming this coding to bachelor's degree = 0 and master's degree = 1 (variable name = degreeE). This is done below and results in a more meaningful constant (intercept). Finally, expE2 represents the subject's years of professional experience. expE2 drops the four most experienced subjects, all with 25 or more years of experience. expE2 also drops those observations dropped by using qptrng5E rather than qptrng. That is, subjects without any quality and productivity training in the past ten years were dropped and an extreme outlier reporting 1514 days of training was dropped. The result of screening out these subjects is a sample more representative of evaluators in training. The regression equation with degree2 simplified by linear transformation as described above is:

Eavg = 21.68 - 6.46 engrdeg + 8.68 degreeE + 0.527 expE2

The 42 subjects used to produce this equation had an average of 8.1 years of work experience (median = 7 years, range = 0 to 21 years). The highest degree completed for most (78.6%) of the subjects was in engineering and 19% had completed a master's degree. The remainder of the subjects had at least a bachelor's degree.

The following equation initially appeared to best predict dimensional accuracy (DAavg) for a sample representative of evaluators in training. This equation was based on data from 54 subjects and yielded an R squared (adjusted) of 0.399 with seven predictors.

$$\text{DAavg} = 19.1 - 0.426\ \text{exp3D} - 5.79\ \text{std} + 0.506\ \text{yrsqcqa} + 4.22\ \text{stloc} - 2.40\ \text{gndr} - 2.74\ \text{size3D} + 3.27\ \text{fed}$$

where exp3D represents the subject's years of professional experience. exp3D dropped the four most experienced subjects, all with 25 or more years of experience. exp3D also dropped the extreme outlier (obs. 29) and three subjects (obs. 19, 47, and 60) originally omitted due to missing values for other variables. This allowed the regression procedure to use the same data as used in stepwise regression. std represents whether or not the subject is a full time student (if yes, value = 1). yrsqcqa represents the number of years the subject had worked in the quality control, quality assurance, or quality improvement function. stloc (state and local government) and fed (federal government) were categories used to describe the subject's employer. gndr represented the subject's gender (female = 1, male = 0). size3D dropped all subjects reporting their employer's number of employees as zero (i.e., those not employed). While a number of subjects were full time students, many were employed[26] and thus not omitted by this variable. Including only those subjects currently employed seemed to greatly increase prediction while also increasing generalizability to those training to be evaluators.

An alternative version of this equation was produced when stloc was withheld from the stepwise regression. This equation was also based on the data from 54 subjects, yielding an R squared (adjusted) of 0.391 while using only six predictors.

[26] Interesting, most full time students who were employed were employed by state or local government. Based on the above prediction equation, a full time student working for state or local government is predicted to be more accurate than other subjects employed by state or local government.

$$\text{DAavg} = 22.8 - 0.421\ \text{exp3D} - 4.64\ \text{std} + 0.532\ \text{yrsqcqa} - 3.07\ \text{size3D} - 2.58\ \text{mfg} - 4.76\ \text{svc}$$

This equation dropped the use of gender (gndr) as a predictor. When stloc was dropped, stepwise regression replaced both stloc and fed with the other two descriptors for the subject's employer: mfg (manufacturing) and svc (service). Interestingly, this new equation predicts a subject working in manufacturing or service will be more accurate than a subject working for the federal, state, or local government. Also interesting, but of questionable significance, is the greater accuracy predicted for those working in service over those working in manufacturing.

Comparison of these two regression equations and their accompanying analyses resulted in noting an observation with an extremely large standard residual. Observation 62 had a standard residual of 3.80 for the first equation and 3.40 for the second equation, clearly high enough to fail tests of residual analysis. Dropping this extreme outlier resulted in a great increase in prediction (R-squared adjusted = 0.538) and the following equation:

$$\text{DAavg} = 25.3 - 0.525\ \text{exp3D} - 6.29\ \text{std} - 2.80\ \text{mfg} - 5.23\ \text{svc} + 0.631\ \text{yrsqcqa} - 3.73\ \text{size3D}$$

The 53 subjects used to produce this final recommended equation had an average of 6.9 years of professional experience (median = 5.5 years, range = 0 to 21 years). The subjects had worked an average of 1.2 years in QC/QA/QI (range = 0 to 15 years). 26.4% of the subjects described themselves as full time students. 41.5% of the subjects were from a large (over 5000 employees) organization. 39.6% of the subjects were from a medium size (between 501 and 5000 employees) organization. The remainder, 18.9%, were from small organizations (1 to 500 employees). The subjects described their employers as follows: manufacturing (15%), service (24.5%), federal government (32%), and state or local government (28%).

The complete regression analyses for the equations presented in this section are shown in Appendix AP: Edited Minitab Session Files for Proposed Regression Equations.

V. Results

This chapter summarizes the key results of the data analyses question by question. Frequent references will be made to tables and figures from Chapter IV: Data Analyses. The focus here is drawing meaning from the data. Where the data were inconclusive, it will be noted and little explanation will be offered. A summary of the results of the analyses is presented in Table 33 at the end of the chapter. The next chapter will integrate results across questions and offer overall conclusions.

Question 0 - For a given sample of untrained evaluators, what is the consistency of their scores?

Question 0 was a broad umbrella question addressed by comparing the two groups' first evaluation scores by item and category, supported by testing hypothesis one (H1) and addressing research questions one through four. Graphically comparing the scores of the control and treatment groups provided an indication of both between and within group consistency (see Figure 3). At first glance, the central location of the scores of the two groups appear to differ noticeably from item to item and category to category. Systematic comparison reveals no patterns and this is later supported by statistical testing. The amount of dispersion as shown by the interquartile range of the boxplots differs widely between groups and between items. For example, compare the boxplots of the items of Category 1.0 - Leadership to the boxplots of the items of Category 7.0 - Customer Focus and Satisfaction. For the items of Category 1.0, the control group's scores consistently show more dispersion than the treatment group's scores. For the items of Category 7.0, the two groups show comparable dispersion on the scores of three items, but the treatment group's scores show more dispersion on the other two items. Even with the averaging effect that occurs when calculating category scores, differences in both location (e.g., Categories 2.0, 5.0, and 7.0) and dispersion (e.g., Categories 1.0, 4.0, and 7.0) of scores appear. Testing hypotheses 1, 2, and 3 showed these differences were not statistically significant and reinforced the expected baseline of no differences between the control and treatment groups.

Hypothesis 1: There will be no difference in scores between the treatment and control groups during the initial evaluation: a) by item and b) by category.

Hypothesis 1 (H1) was tested by using a t-test to compare the mean of the control group's scores to the mean of the treatment group's scores for each item and category. The results in Table 2 show there were no differences between the mean scores of the two groups on any item or category. Given that there were twenty-four independent t-tests conducted for the twenty-four items, one or more significant differences could have been expected due to random chance when alpha = 0.05. A test of the scores prior to screening (i.e., using all available data) produced the same result: no significant differences. Since category scores are simply an average of subsets of the item scores, it is not surprising that the tests of the category scores mirrored those of the corresponding item scores. These results established the expected baseline of no difference in mean scores between the control and treatment groups.

Question 1 - How much agreement is there among evaluators on the score of an item?

The relative agreement of the evaluators within each group was illustrated by the magnitude of the interquartile range and whiskers of each item score boxplot in Figure 3. As discussed under Question 0, this agreement or dispersion varied widely. For example, the control group's interquartile range (IQR) for item 4.2 was 40 points (on a 100 point scale) but the same group's IQR for item 4.3 was only 10 points. The whiskers on each boxplot show the range of scores for that group on that item, excluding outliers. Whiskers with ranges as low as 10 points and as high as 80 points were seen; however, these extremes were not seen on items of the same category. Without a basis for comparison, these illustrations of within group agreement are primarily descriptive. They do provide a baseline of comparison for future research. Item score agreement was further examined by testing Hypothesis 2, which used variance as the measure of agreement.

Hypothesis 2: For each item, there will be no difference in score variances between the treatment and control groups.

Hypothesis 2 (H2) used an F-test to compare the variance of the control group's scores to the variance of the treatment group's scores for each item. The results in Table 3 show there were no differences between the variances of groups' scores on any item. The test of the variances for Item 4.3, Employee Education, Training, and Development, did yield a p-value of 0.0503, but this was deemed inconsequential. With twenty-four independent F-tests using an alpha of 0.05, a single rejection of the null hypothesis is likely to occur by random chance. These results continued the establishment of a baseline of no differences between the scores (in this case, item score variances) of the control and treatment groups.

Question 2 - How much agreement is there among evaluators on the score of a category?

The relative agreement of the evaluators within each group was illustrated by the magnitude of the interquartile range and whiskers of the category score boxplots in Figure 3. The effect of averaging item scores, or regression toward the mean, can be seen in the increased consistency of the plots of the category scores. In most cases, the interquartile range of the category scores was smaller than the simple average of the interquartile ranges of the corresponding item scores. Whiskers with ranges as low as 11.5 points and as high as 60 points were seen for the control group. Whiskers from the treatment group's category score plots ranged from 20 to 60 points. As with the within group agreement of item scores, this information is primarily descriptive and provides a future basis of comparison.

The discussion of Question 0 mentioned some apparent differences in level of agreement between the control group and treatment group. That is, the IQR of the two groups' scores look quite different for Categories 1.0, 4.0, and 7.0. For example, the control group's IQR for Category 1.0 spans 30 points, while the treatment group's IQR for the same category spans only 13.3 points. On the other hand, the IQR of the two groups' scores appear almost equal for Categories 2.0, 3.0, 5.0, and 6.0. The statistical significance of the differences in category score agreement, as measured by score variances, was tested by Hypothesis 3.

Hypothesis 3: For each category, there will be no difference in score variances between the treatment and control groups.

Hypothesis 3 (H3) used an F-test to compare the variance of the control group's scores to the variance of the treatment group's scores for each category. The results in Table 4 show there were no differences between the variances of the groups' scores on any category. The lowest p-value from testing any of the categories was 0.276, providing no evidence of a difference between score variances. This is not surprising given the results of testing H2 and adds to the expected baseline of no differences between the scores (in this case, category score variances) of the control and treatment groups.

Question 3 - How consistent is the within-item variation of the evaluator scores across all the items of a category?

The consistency of evaluator agreement across the items of a category was illustrated by comparing the magnitude of the interquartile range and whiskers of the item score box plots in Figure 4. The box plots in Figure 4 were constructed using the combined data from the control and treatment groups' first evaluation. This basically doubled the sample size for each plot and was expected to produce a more reliable distribution. The boxplots for all the items of each category were plotted on a single chart to facilitate comparison.

The boxplots for Categories 1.0 and 3.0, and perhaps 5.0 and 6.0 showed the expected effect. That is, little difference was expected between the dispersions of item scores across a particular category. The plot of Item 1.1's scores had one notably shorter whisker than Items 1.2 and 1.3, but the three outliers (i.e., the asterisks in the plot) on the short side of Item 1.1's plot imply a longer whisker is plausible. The plots of the item scores for Category 3.0 showed only a small difference in whisker length. The plots of the item scores for Category 5.0 show small differences in whisker lengths and IQR. This may simply reflect random error in the Category 5.0 item scores. Two of the three plots of item scores for Category 6.0 were almost identical. The plot of Item 6.2's scores had both a wider IQR and a longer lower whisker than Items 6.1 and 6.3. Item 6.1 had a single outlier and Item 6.3 had two outliers on the lower side, which may be too few points to imply a longer whisker. The plots of all three items of Category 6.0 show relatively less dispersion than the items of any of the other categories.

Some notable differences in item score dispersion appeared in the plots of Categories 2.0, 4.0, and 7.0. Items 2.1 and 2.3 had very similar boxplots, but Item 2.2 had both a wider IQR and longer whiskers. Review of the content of these items offers no obvious reason for this difference. Wider differences appeared in the plots for Category 4.0. Items 4.1 and 4.2 appear to have more dispersion than Items 4.3 and 4.4. Because these items deal with a relatively subjective area, human resource management, it is not a surprise they have wide dispersions. Wide dispersion was seen in the whiskers of the boxplot of Item 4.1 and the IQR of Item 4.2. Review of the content of these items show that Items 4.3 and 4.4 deal with more easily quantifiable human resource issues: employee education, training, and development (4.3) and employee well-being and satisfaction (4.4). Items 4.1 and 4.2 deal with the more qualitative issues of human resource planning and evaluation and high performance work systems, respectively. The plots of item scores for Category 7.0 produced an interesting pattern. The plots of Items 7.1 and 7.2 were quite similar, with relatively narrow IQRs and short whiskers[1]. The plots of Items 7.3, 7.4, and 7.5 had notably wider IQRs and longer whiskers. It appeared that perhaps two different constructs were being measured by this set of items. Review of the item's content showed

[1] Although Item 7.1's plot did have three outliers on the lower side.

this to be the case. Items 7.1 and 7.2 are both related to issues of customer focus. Items 7.3, 7.4, and 7.5 are all related to measures of customer satisfaction.

Question 4 - How consistent is the within-category variation of the evaluator scores across all seven categories?

The consistency of the dispersion of category scores was illustrated by comparing the magnitude of the interquartile range and whiskers of the category score boxplots in Figure 5. The overall range of scores for Category 6.0 is noticeably shorter than the other categories. The range of scores for Category 7.0 is slightly longer, with all the others showing much longer whiskers and Categories 2.0 and 3.0 showing wider IQRs as well. This is likely due to more than random chance. The content of Category 6.0 is arguably the most quantitative of any category and it is scored using only the results scoring guidelines. Category 7.0 is somewhat quantitative, but is scored using both the approach/deployment and results scoring guidelines. Category 5.0 is also somewhat quantitative, but is scored using only the approach/deployment scoring guidelines. The other four categories are scored using only the approach/deployment scoring guidelines. The subjectivity of the content and the guideline (scale) used are both potential explanations for this pattern of dispersion. The statistical significance of these differences, as measured by within-category score variances, were tested by Hypothesis 4.

Hypothesis 4: There will be a difference in score variances between categories.

Hypothesis 4 used Hartley's F_{max} test (Ott, 1984) to test for homogeneity of variance across the seven categories. The results of the test were inconclusive. That is, the test statistic fell between two critical values in Pearson and Hartley's table of values for F_{max}. The conservative approach would be to rule the test results as insignificant, since the test statistic is smaller than the critical value with the next lowest degrees of freedom. Instead, this result was used as justification for further analysis.

Pairwise comparisons of the variances of the seven categories provided evidence of a more meaningful difference. No pairwise comparison yielded a significant difference when the ultraconservative Bonferroni adjustment was used. With the Bonferroni adjustment, a p-value less than or equal to 0.0024 was required to reject the null hypothesis. While this controlled for a type-I error, it also greatly reduced the power of the test. Seven out of twenty-one pairwise comparisons resulted in p-values less than the experiment-wise error rate of 0.05. This result was very unlikely[2] as a result of random chance. The p-values in Table 7 show a clear pattern of differences between score variances of the seven categories. The score variances of Categories 6.0 and 7.0 appear to be different from the score variances of the other five categories. Comparing the variance of Category 6.0's scores to the variances of the other six categories resulted in p-values less than 0.05 in four of the six comparisons and a p-value of 0.052 in the fifth comparison. Only when compared to the score variance of Category 7.0 was a large p-value (p = 0.786) produced. Comparing the variance of Category 7.0's scores to the variances of the other six categories resulted in p-values less than 0.05 in three of the six comparisons. One of the non-significant comparisons was with Category 6.0 (already mentioned, p = 0.786). The remaining two non-significant comparisons produced p-values of 0.076 and 0.110. This quantitative evidence, combined with the previously described differences in subjectivity of content and scale, support the conclusion that the scores of Categories 6.0 and 7.0 have less dispersion than the other five categories.

Question 5 - How accurate are the evaluators' scores for each category?

The accuracy of the evaluators' scores was illustrated in Figure 6 and summarized by the mean and median accuracy indices in Table 8. Figure 6 superimposed the experts' scores over the boxplots from Figure 3 (i.e., boxplots of each group's scores on each item and category). A clear pattern of poor elevation accuracy emerges from reviewing Figure

[2] For a binomial distribution with a 0.05 chance of success, the probability of achieving seven successes in twenty-one trials is 0.0000.

6. In every case, the untrained evaluators' median scores are higher than the experts' scores. For the majority of items and categories, the untrained evaluators' first quartile scores are higher than the experts' scores. This implies a leniency effect on the part of the untrained evaluators. This relatively large and consistent difference between the experts' scores and the evaluators' scores is also reflected in the mean and median elevation indices in Table 8. It appears that the untrained evaluators have some difficulty assigning scores based on the descriptive anchors of the scoring guidelines. The pattern of assigning scores 10, 20, or even 30 points higher than the scores assigned by the experts is unmistakable.

Figure 6 does not provide a clear indication of dimensional accuracy. The one-to-one correspondence or tracking inherent in dimensional accuracy is lost when comparing a set of distributions to a set of scores. The DA indices in Table 8 indicate moderate dimensional accuracy for the untrained evaluators. Graphic examples illustrate the difficulty of interpreting DA from a set of distributions and assist with the interpretation of the indices in Table 8. Figures 14 and 15 show how the control group's mean item scores on Category 7.0 are more accurate, in terms of DA, than the mean DA of the control group's individual scores on Category 7.0. The control group's mean item scores on Category 7.0 yield a DA index of 3.8, while the control group's mean dimensional accuracy on Category 7.0 was 11.0. This is not surprising, since the mean item scores should be better estimators of the true scores (i.e., expert scores) than the item scores of an individual. The control group's mean item scores shown in Figure 14 are an example of good dimensional accuracy. Notice how changes in score from item to item nearly parallel the changes in the experts' scores. Comparing Figures 14 and 15 gives an indication of the difference of between a DA index of 3.8 and approximately 11.0. Figure 15 compares the scores of a subject with near average[3] dimensional accuracy to the experts' scores on Category 7.0. Notice how the scores of the subject with average DA parallel the experts' scores on Items 7.1 and 7.2, then drop 30 points for Item 7.3 while the experts' score drops only 15 points. While the experts' score on Item 7.4 is 15 points

[3] The subject is from the control group and is described in terms of the control group's average DA.

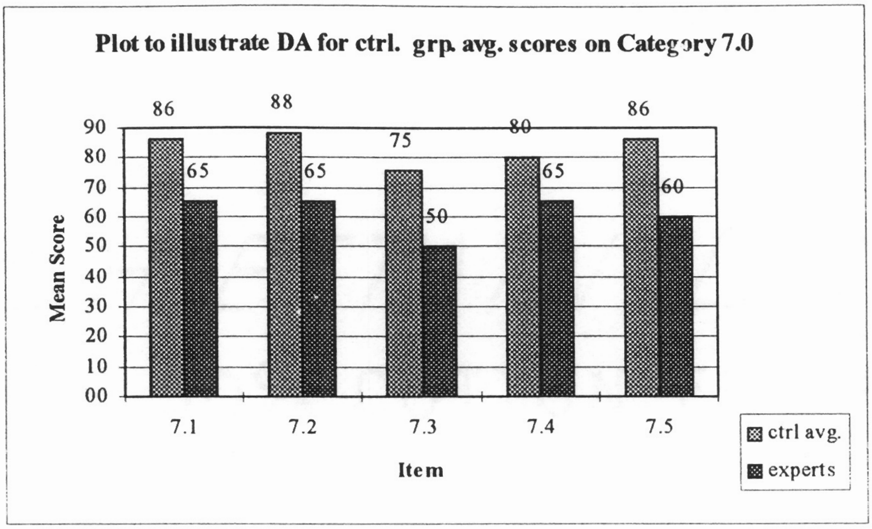


Figure 14.

Comparison of the control group's mean scores on Category 7.0 to the experts' scores (DA index = 3.8, elevation index = 22.2).

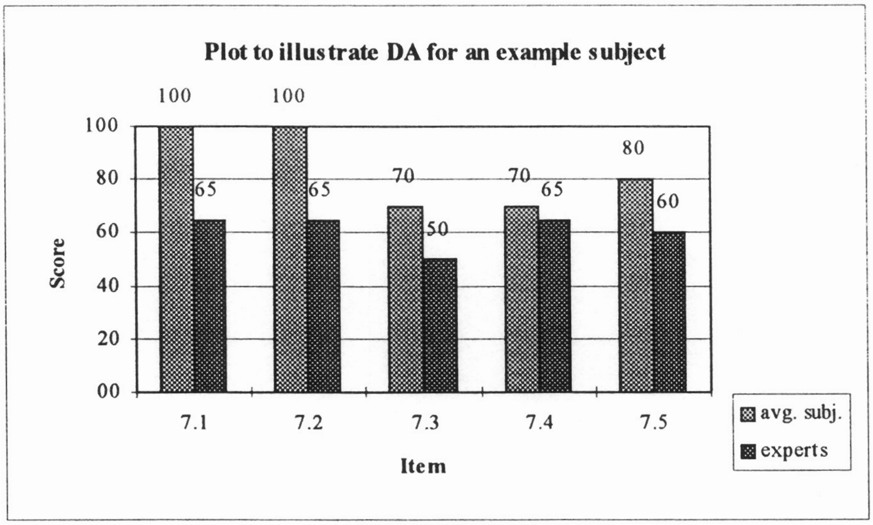


Figure 15.

Comparison of experts' scores to the scores of a subject with near average dimensional accuracy on Category 7.0 (this subject's DA index = 11.2, elevation index = 23.0).

higher than their score on Item 7.3, the average subject gave the two items equal scores. Finally, the experts scored Item 7.5 as 5 points lower than Item 7.4, but the average subject scored Item 7.5 as 10 points higher than Item 7.4. The dimensional accuracy of the subject's scores in Figure 15 is fairly representative of the DA of the typical subject's scores. The mean DA for each category ranged from 7.82 to 18.11 and the median DA for each category ranged from 5.0 to 17.5. Figure 16 compares the scores of a subject with poor dimensional accuracy to the experts' scores on Category 7.0. Notice how poorly the subject's scores parallel those of the experts.

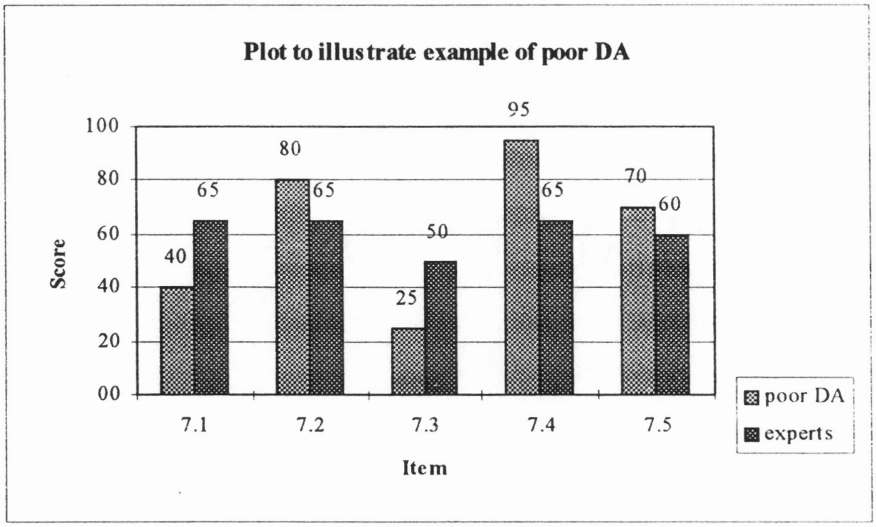


Figure 16.

Comparison of experts' scores to the scores of a subject with poor dimensional accuracy on Category 7.0 (this subject's DA index = 22.2, elevation index = 1.0).

The above discussion of the evaluators' scores was primarily descriptive and provides a baseline of comparison for future measures of evaluator accuracy. The data did indicate a leniency effect. This is not surprising since exposure to organizational practices at the Baldrige contender level of excellence is more likely to be the exception than the rule.

Hypothesis 5 compared the central location (mean and median) of the control and treatment groups' accuracy indices for each category. Both classical (t-test) and rank based (Mann-Whitney test) tests were used and yielded the same results (see Table 9). These results completed the establishment of a baseline of no differences between the scores (in this case, both elevation and dimensional accuracy) of the control and treatment groups. This was important since elevation and dimensional accuracy were not believed to have been used with these types of scores before. Plotting the distributions of subjects' accuracy indices yielded a variety of patterns, but the agreement of the classical and rank based tests imply these statistics may be useful for analysis of evaluator scores.

The results of the content analysis of the evaluator's first evaluation comments on strengths, areas for improvement, and site visit issues were summarized in Table 10. No statistical tests were performed on these data. Table 10 presents the data and summary statistics for the combined groups; however, Appendix AS presents the data and summary statistics by group. Some patterns emerged when these data were reviewed in a gestalt fashion. The overall percent of the experts' comments identified (%Experts) for each item was the most stable indicator. That is, for a given item the control and treatment groups' %Hits might vary widely due to one or more subjects submitting only one comment that was an obvious match or hit. This subject's 100% hits has a large influence on the group's average %Hits. This same situation had less impact on the group's %Experts. %Experts was also more stable among the items of a category. The largest range observed for %Experts was for Category 7.0. Overall, the subjects identified a low of 6% of the experts' comments on Item 7.5 and a high of 18% on Item 7.1. The %Experts on Item 7.5 was the lowest observed and might have been due to fatigue[4]. Item 7.5 was likely the last item scored by all the subjects scoring Category 7.0. Item 1.1 had the highest overall %Experts, with the subjects identifying an average of 30% of the experts' comments. Other items from Category 1.0 also had relatively high (25% or more)

[4] Item 7.5 had one of the lowest number of average comments submitted.

%Experts and this further supports the fatigue explanation for Item 7.5. The items of Category 1.0 were likely the first items scored by all the subjects scoring Category 1.0.

The low overall %Experts values were partially due to the limited number of subjects identifying site visit issues. For most items, less than half the subjects identified site visit issues. Most subjects identified areas for improvement, but they usually identified a larger number of strengths. This is reflected in the %Experts for strengths. The average percent of "true" strengths identified by the subjects ranged from a low of 14% on Item 7.5 to a high of 59% on Item 1.3. Within each of the seven categories, the item for which the most true strengths were identified had a %Experts for strengths of at least 38%.

Using averages of the %Hits and %Experts statistics across items or categories may limit validity, but it provides an illustrative example. The average subject on the typical item identified 36% of the strengths identified by the experts and 17% of the overall comments of the experts. 35% of this average subject's comments were hits when identifying strengths and 28% were hits overall.

For a given sample of untrained evaluators, will evaluator training change the consistency of their scores?

This broad umbrella question introduced the post-treatment analysis. This question was addressed by comparing the two groups' first and second evaluation scores by item and category, supported by testing hypothesis six (H6) and addressing research questions six through nine. Graphically comparing the pre- and post-treatment scores of the control and treatment groups provided an indication of changes in consistency (see the box plots in Figure 7). A pattern emerges in the central tendency of the scores as shown by the medians of the box plots. The median of the treatment group's second evaluation scores appears to be equal or less than the median of their first evaluation scores. On the other hand, the median of the control group's second evaluation scores appears to increase or decrease with comparable frequency. This trend appears for every item and category. A review of the mean scores in the descriptive statistics of Appendix AF supports[5] the trend shown in the box plots. This suggests that training the evaluators may have reduced the leniency effect observed in the pre-treatment scores. Testing hypothesis 6 further highlighted these patterns and showed that the changes in the treatment group's mean scores were sometimes statistically significant. No clear patterns were seen in the dispersions of the scores as shown by the interquartile ranges and whiskers in the box plots.

Hypothesis 6: There will be a difference in scores between the control and treatment groups during the second evaluation: a) by item, and b) by category.

The reason for the hypothesized difference in H6 is the training intervention the treatment group received between the first and second evaluations. To demonstrate that the difference was not simply a function of learning from the first evaluation, the factor of time was also considered. Thus, hypothesis 6 (H6) was tested using a two-way (group x time) analysis of variance (ANOVA). Evidence of an interaction or main effect was followed up with a test of simple effects using a t-test. The complete results of these tests were shown in Tables 11 (items) and 12 (categories). A summary of the significant results is provided in Table 32. No significant effects were observed for the items of Category 1.0. Evidence of main effects for each of the items of Category 2.0 led to tests for simple effects. The only significant ($p = 0.042$) result was a simple time effect found on Item 2.1 for the evaluators receiving training. Evidence of a main time effect on Item 4.1 led to the discovery of a significant ($p = 0.018$) simple time effect for the evaluators that did not receive training. This was the only significant time effect observed for the evaluators who did not receive training. Evidence of effects were observed for three of the four items of Category 5.0. No effects were observed for Item 5.1 and evidence of an interaction for

[5] The treatment group's mean scores on five items showed a slight increase from the first to second evaluation, typically two or three points, where the median scores had shown no increase. In each case, the control group's mean scores on that item showed an even larger increase from the first to second evaluation.

Table 32. Summary of Significant Results from Testing Hypothesis 6 by Item.

Item	Interaction	Main Effects		Simple Effects			
	Effect	Group	Time	Group		Time	
				Pre-	Post-	Control	Treatment
2.1	0.656	0.462	0.064*			0.41	0.042**
2.2	0.530	0.071*	0.163	0.38	0.078		
2.3	0.449	0.291	0.046**			0.066	0.36
4.1	0.246	0.244	0.084*			0.018**	0.71
4.3	0.953	0.071*	0.215	0.24	0.13		
5.2	0.057*			0.20	0.16	0.063	0.61
5.3	0.053*			0.51	0.035**	0.18	0.18
5.4	0.013**			0.27	0.02**	0.063	0.12
6.1	0.087*			0.95	0.043**	0.15	0.0006**
6.2	0.032**			0.69	0.0003**	0.40	0.0009**
6.3	0.103	0.044**	0.054*	0.77	0.016**	0.81	0.022**
7.3	0.609	0.042**	0.244	0.30	0.03**		
7.4	0.467	0.024**	0.422	0.14	0.11		
7.5	0.564	0.012**	0.244	0.085	0.085		

* = mild evidence of an interaction or main effect, $p \leq 0.10$
** = evidence of significant effect, $p \leq 0.05$

Item 5.2 was inconclusive[6]. Evidence of an interaction for Item 5.3 led to the discovery of only one significant simple effect: a simple group effect (p = 0.035) for the post-treatment comparison. A significant (p = 0.013) interaction was observed for Item 5.4. Examination of simple effects found a significant (p = 0.02) simple group effect for the post-treatment comparison. Two of the four items in Category 5.0 showed a significant post-treatment difference between groups, while no significant differences were observed pre-treatment. Significant effects were observed for each of the three items in Category 6.0. Evidence of an interaction for Item 6.1 led to the discovery of a significant (p = 0.043) simple post-treatment group effect and a very significant (p = 0.0006) simple time effect for the

[6] Examination of the means for Item 5.2 showed a classic interaction pattern. The mean score of the control group increased from 62 to 84, while the mean score of the treatment group decreased from 75 to 71. None of these changes were large enough to produce a significant simple effects result.

evaluators receiving training. A significant (p = 0.032) interaction on Item 6.2 led to the discovery of similar simple effects. For Item 6.2, a very significant (p = 0.0003) simple post-treatment group effect and a very significant (p = 0.0009) simple time effect for the evaluators receiving training were observed. This directly supports the results of the tests of the scores from Item 6.1. Little evidence of an interaction was seen for Item 6.3, yet a significant (p = 0.044) main group effect and evidence (p = 0.054) of a main time effect were observed. Examination of simple effects for Item 6.3 produced results consistent with those of Items 6.1 and 6.2. For Item 6.3, a significant (p = 0.016) simple post-treatment group effect and a significant (p = 0.022) simple time effect for the evaluators receiving training were observed. The results of the tests for Items 6.1, 6.2, and 6.3 suggest the scores of the groups were different only after training and that the only group whose score changed was that receiving training. Significant main group effects were observed for Items 7.3, 7.4, and 7.5. This could be interpreted to suggest an underlying difference between the groups. Examination of the mean scores shows the treatment group's mean scores are 7 to 9 points lower on the first evaluations[7], and this difference increases for the second evaluations. Averaging these differences over time resulted in the significant main effect. Examination of simple effects found only one significant effect, a simple post-treatment group effect (p = 0.03) for Item 7.3.

Testing the scores by category (see Table 12), produced a condensed version of the tests by item. This was expected, since category scores are the mean of the item scores in that category. Significant interactions were observed for Category 5.0 (p = 0.036) and Category 6.0 (p = 0.028). Examination of simple effects for Category 6.0 found a very significant (p = 0.0038) simple post-treatment group effect and a very significant (p = 0.0005) simple time effect for the evaluators receiving training. No other significant simple effects were observed.

A clear pattern emerged from the ANOVA and t-tests conducted for H6. Evidence of differences in mean scores was clustered among the items and categories with

[7] These differences were shown to be insignificant in testing H1.

relatively quantitative content. The results clearly implied that the training intervention made a difference in the scores of the evaluators on the items of Category 6.0. The content of Category 6.0, Business Results, is the most quantitative of the seven categories and the items of Category 6.0 are scored using only the results guideline. All other categories use the approach/deployment guideline or a combination of the two. Category 5.0, Process Management, is somewhat quantitative and may appear less subjective than Categories 1.0 to 4.0 appear to a group of evaluators dominated by engineers. Category 5.0 is scored using only the approach/deployment guideline. The results regarding Category 5.0 scores were less conclusive than those for Category 6.0, but did show a difference between the trained and untrained evaluators on Items 5.3 and 5.4. Category 7.0, Customer Focus and Satisfaction, is also somewhat quantitative. Category 7.0 is the only category that uses a combination of the scoring guidelines. Items 7.1-7.3 use the approach/deployment guideline and Items 7.4-7.5 use the results guideline. The significant group main effects for Items 7.3-7.5 are somewhat inconclusive. Since no significant differences were found for the pre-treatment simple group effects, it could have been the influence of the post-treatment simple group effects that produced these main effects. However, only one of the post-treatment simple group effects was significant, Item 7.3. The difference in mean scores for Items 7.3-7.5 was greater post-treatment than pre-treatment, but this was inadequate for drawing meaningful conclusions. The significant simple time effect for untrained evaluators observed on Item 4.1 and significant simple time effect for trained evaluators on Item 2.1 were also deemed inadequate for meaningful conclusions.

Question 6 - Did agreement among evaluators on the score of an item change (improve) due to evaluator training?

The relative agreement among evaluators on the score of an item was illustrated by comparing the magnitude of the interquartile range and whiskers of the item score box plots in Figure 7. The first evaluation (pre-training) and second evaluation (post-training) distributions were compared for each group, followed by a comparison of the second evaluation distributions of the two groups. No discernible patterns emerged from this graphical analysis. Dispersion of second evaluation item scores seemed to increase and decrease randomly, like that observed on the first evaluation scores. The magnitude of item score dispersion appeared no different for the second evaluation than the first. This was not the expected result. The treatment group's trend of decreasing median scores from the first to second evaluation might be expected to increase the overall range of their scores. Comparing the range in terms of the box plot whiskers did not support this explanation. There does not appear to be a pattern of increasing or decreasing agreement among the evaluators on the scores of items. Hypothesis 7 (H7) tested the statistical significance of any changes in evaluator agreement from first to second evaluation, using item score variance as the measure of agreement. Hypothesis 7b tested for differences in evaluator agreement between the control and treatment groups' second evaluation item scores.

Hypothesis 7: Item score variances will be smaller for second evaluation scores than for first evaluation scores.

H7 was tested by using an F-test to compare the variance of the treatment group's second evaluation item scores to the variance of the treatment group's first evaluation item scores. A one-sided test was used to see if the second evaluation score variances were smaller than the first evaluation score variances. The results shown in Table 13 provide no evidence that the training intervention improved evaluator agreement on the score of an item. H7 was also tested to compare the variance of the control group's second evaluation scores to the variance of the control group's first evaluation scores. Again, a one-sided test was used to see if the second evaluation score variances were smaller than the first evaluation score variances. Very significant differences ($p<0.01$) were observed for three (Items 1.3, 4.1, and 4.4) out of twenty-four items. These differences appear to be random. The control group's score variance on Item 1.3 was the largest variance observed during their first evaluation. Thus, the difference in score variance on Item 1.3 from first to second evaluation could be due to regression toward the mean. Examination of the control group's second evaluation score variance on Item 1.3 shows it to be of typical size, not unusually small. The control group's second evaluation score variance on Item 4.1 was the smallest variance observed for either group on either evaluation. Comparing this small variance to many of the first evaluation score variances would have produced a significant result. Interestingly, both groups' score variance decreased noticeably on Item 4.1 from the first to second evaluation. An outlier in the treatment group's second evaluation scores (see Figure 7) prevented a significant result for the treatment group. The control group's large first evaluation score variance on Item 4.4 can be explained by the presence of a single outlier (see Figure 7). Without this outlier, no significant difference would have been found between the first and second evaluations. Thus, the decreased variances observed for the control group on Items 1.3 and 4.4 appear to be a function of randomly large first evaluation score variances.

Testing Hypothesis 7b produced no significant results. The treatment group's second evaluation item score variances were not significantly smaller than the control group's second evaluation item score variances. Thus, no evidence was seen that the training intervention reduced subsequent item score variances, thereby increasing evaluator agreement on the scores of items.

Question 7 - Did agreement among evaluators on the score of a category change (improve) due to evaluator training?

The relative agreement among evaluators on the score of a category was illustrated by comparing the magnitude of the interquartile range and whiskers of the category score box plots in Figure 7. A faint pattern emerged when the trends of the box plots were compared to the descriptive statistics for the category scores. Whenever the central tendency of the control group's category scores decreased noticeably (from first to second evaluation), the variance of their category scores increased. Two of three times the central tendency of the control group's category scores increased noticeably, the variance of their scores decreased. The one exception, Category 5.0, showed a large increase in the control group's mean and median scores, but little change in variance. This trend might be explained by the leniency effect observed among the untrained evaluators. When the central tendency of the scores is near the upper bound, an increase in the central tendency is likely to produce a decrease in variance and a decrease in central tendency is likely to produce an increase in variance. This is a result of restricted range near the boundaries. The exception can be explained by the fact that the control group's first evaluation mean score (71.1) on Category 5.0 was the lowest mean score observed for any category during the first evaluation. Thus, the problem of range restriction on the upper side was less for the control group on Category 5.0 than any of the other categories or for the treatment group on any category. This same trend was not observed for the treatment group. The central tendency of the treatment group's scores decreased noticeably[8] in five of seven categories. The variance of the treatment group's scores decreased for two of these five categories and increased for the other three categories. H8 tested the statistical significance of any changes in evaluator agreement from first to second evaluation, using category score variance as the measure of agreement. H8b tested for differences in evaluator agreement on category scores between the control and treatment groups' second evaluation category scores.

Hypothesis 8: Category score variances will be smaller for second evaluation scores than for first evaluation scores.

H8 was tested by using an F-test to compare the variance of the treatment group's second evaluation category scores to the variance of the treatment group's first evaluation category scores. A one-sided test was used to see if the second evaluation score variances were smaller than the first evaluation score variances. The results shown in Table 16

[8] The central tendency of the treatment group's scores on the remaining two categories was approximately the same from the first to the second evaluation.

provide no evidence that the training intervention improved evaluator agreement on the score of a category. H8 was also tested to compare the variance of the control group's second evaluation category scores to the variance of the control group's first evaluation category scores. Again, a one-sided test was used to see if the second evaluation score variances were smaller than the first evaluation score variances. Significant differences ($p<0.05$) were observed for two (Categories 1.0 and 4.0) out of seven categories. These results may partially[9] reflect the results found when testing item score variances for H7. A more likely explanation is that observed during the graphical analysis. For both Categories 1.0 and 4.0, the central tendency of the control group's scores increased from the first to second evaluation. Thus, this increased agreement (reduction in variance) among the control group is likely a function of continued leniency effect coupled with upper range restriction. Given the continued leniency effect, the control group's category scores are expected to tend toward the upper bound. A random increase in the central tendency of the control group's category scores could easily magnify a random reduction in variance due to the upper range restriction.

Testing H8b produced no significant results. The treatment group's second evaluation category score variances were not significantly smaller than the control group's second evaluation category score variances. No evidence was seen that the training intervention reduced the subsequent category score variances, thereby increasing evaluator agreement on the scores of items. It could be that the range restriction caused by the leniency effect masked any decrease in score variance due to the training intervention. That is, in three out of four cells of this experimental design, the evaluators were untrained and believed to suffer from a leniency effect which compressed score variances. Only the trained evaluators (i.e., treatment group second evaluation) appeared free of the leniency effect and thus displayed uncompressed score variances.

[9] The variance of average (category) scores is not the same as the average variance of individual (item) scores.

Question 8 - Did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?

The within-item score variation (dispersion) across the items of a category was illustrated by comparing the magnitude of the interquartile range and whiskers of the item score box plots in Figure 8. The box plots in Figure 8 were constructed using the score data from the treatment group's first and second evaluation. These graphics were supplemented by reviewing the item score standard deviations from the descriptive statistics of Appendix AF. Since the treatment group received training between their first and second evaluations, the dispersion of their scores was expected to decrease.

The H7 tests that compared first to second evaluation item score variances showed no statistical differences for the treatment group. These tests were made on an item-by-item basis and did not look for patterns of change across the items of each category. Review of the box plots did not reveal any notable patterns, with the possible exception of Category 4.0. In fact, every category except Category 4.0 had items where the dispersion of scores increased and items where the dispersion of scores decreased. For Category 4.0, the dispersion appeared to decrease for every item. For Items 4.1, 4.3, and 4.4, both the IQR and whiskers were smaller for the second evaluation scores. Item 4.2 displayed a large decrease in IQR, but an increase in the length of the whiskers. Examination of standard deviations showed a decrease in standard deviation for all four items. Results for the control group were very similar. All categories except category 4.0 displayed random patterns of increasing and decreasing item score dispersion from the first to the second evaluation. For the control group, Items 4.1, 4.2, and 4.4 displayed decreases in item score dispersion. The differences for Items 4.1 and 4.4 were statistically significant (see H7). Item 4.3 displayed somewhat of an increase in score dispersion, but examination of standard deviations showed this to be a relatively small increase. Such a pattern of decreasing score dispersion for the items of Category 4.0 might be due to the inherent subjectivity of the content of Category 4.0. The content of Category 4.0, Human Resource Development and Management, is arguably the most subjective of the seven categories. Such subjectivity could be expected to result in widely dispersed scores for the first evaluation. If this is true, a training intervention or simply the experience of a prior evaluation could be expected to reduce this relatively large dispersion. The data do not support this line of reasoning. The control group's first evaluation scores did not show relatively wide dispersion on Items 4.3 and 4.4. Thus, no changes were seen in within-item variation of evaluator scores across ail the items of any of the seven categories due to training. In fact, no changes were seen for the trained or untrained evaluators. Perhaps a more in-depth training intervention is required to effect within-item variation.

Question 9 - Did within-category variation of the evaluator scores across all seven categories change (decrease) due to evaluator training?

The consistency of the dispersion of category scores before and after training was illustrated by comparing the interquartile ranges and whiskers of the box plots in Figure 9 with those in Figure 11. While the downward shift of the medians stands out, no pattern of change is seen in the dispersion of the scores from the first to the second evaluations. Figures 10 and 12 show the dispersion of the untrained evaluators' first and second evaluation category scores. Again, no pattern of change is seen. Graphical comparison of the trained and untrained evaluators' second evaluation scores is also inconclusive. During the first evaluation, the combined scores of Category 6.0 had a noticeably shorter range than the scores of the other categories. This tighter dispersion was not seen in the second evaluation scores of Category 6.0. The statistical significance of differences in within-category score variances was tested by hypothesis 9 (H9).

Hypothesis 9: There will be a difference in score variances between categories for both the control and treatment groups.

Hypothesis 9 used Hartley's F_{max} test (Ott, 1984) to test for homogeneity of variance across the seven categories. The results of the tests showed no difference for the category score variances of the trained evaluators and a strongly significant difference

($p < 0.01$) for the category score variances of the untrained evaluators. Although a difference was hypothesized for the category score variances of the trained evaluators, it was believed that training might reduce this difference below statistical significance. This appears to be supported. Even if the outliers in the trained evaluators' scores on Category 2.0 and Category 4.0 were removed (see Figure 9), the difference in category score variances would still not be statistically significant.

A significant difference was expected for the untrained evaluators, like that seen for the combined (pre-training) first evaluation category scores. The data provided strong evidence of this difference. Even if the outlier in the untrained evaluators' scores on Category 3.0 were removed (see Figure 10), the variance of Category 2.0 scores is large enough to be significantly different than the smallest category score variance (Category 4.0). The relatively small category score variances observed for the untrained evaluators (e.g., on Categories 4.0 and 1.0) may be explained by range restriction as their median scores approached the upper bound. The three categories with relatively small score variance (1.0, 4.0, and 7.0) have noticeably higher medians than the other categories, with the exception of Category 5.0. This does not explain the relatively large[10] category score variances for the untrained evaluators on Categories 2.0 and 3.0. Pairwise comparisons of the untrained evaluators' category score variances reinforced the statistical significance of these differences, but provide no additional explanation. Perhaps the large variances of Categories 2.0 and 3.0 for the untrained evaluators and Categories 1.0 and 3.0 for the trained evaluators were simply a function of small sample size ($n_{avg.} < 10$). The examination of first evaluation category score variances (see Q4) used combined group data, resulting in an average sample size more than twice as large ($n_{avg.} > 22$) as that used for Q9. The pattern of decreasing category score variance in the first evaluation corresponded to increasingly quantitative category content. This expected pattern was not observed in the second evaluation category score variances.

[10] The variance of Categories 2.0 and 3.0 (414 and 594, respectively) were much larger than any of the variances observed from the combined 1st evaluation category scores (largest was Category 3.0 at 303).

Question 10 - Did the accuracy of the evaluators' scores change (improve) due to evaluator training?

The change in accuracy of the evaluators' scores is illustrated in Figure 13 and Appendix AI. The box plots in Figure 13 clearly show a trend of the trained evaluators' median scores moving closer to the experts' scores while the untrained evaluators' scores show a rather random pattern of change. Prior to the training intervention (i.e., during the first evaluation), both group's median item scores were higher than the experts' scores on all twenty-four items. The control group's median scores on the second evaluation were also higher than the experts' scores on all twenty-four items. The treatment group's median scores on the second evaluation appear to match the experts' score on three items (2.2, 5.4, and 6.3) and were actually lower than the experts' scores on one item (2.1). This resulted in the treatment group's median score on Category 2.0 being slightly lower than that of the experts' score on Category 2.0. Based on these graphical comparisons, the training intervention appears to have reduced the leniency effect seen in the pre-training evaluations. This graphical interpretation is supported by the results of hypothesis 10.

The results[11] of H10 provided evidence that the training intervention had a significant effect on elevation accuracy. While only one significant interaction effect was seen for elevation accuracy, the test for simple group effects showed the trained evaluators were significantly more accurate after training than the untrained evaluators in five out of seven categories (Categories 3.0 to 7.0). Mild evidence that the trained evaluators were more accurate was seen for one of the two remaining categories, Category 2.0. From pre- to post-training, the trained evaluators' mean elevation accuracy improved for all seven categories. This improvement was statistically significant (p = 0.023) for Category 2.0

[11] This discussion is based on the results of the two-way ANOVA and accompanying t-tests. The results of the Friedman-type rank test and the accompanying Mann-Whitney tests concurred with the classical tests. This suggests that the classical tests were robust enough to accommodate any violation of the normality assumption; therefore, the results of the classical tests were used.

and for Category 6.0 (p = 0.001). Mild evidence of a significant difference was seen for Category 3.0 as well. Comparison of the untrained evaluators' first to second evaluation mean elevation accuracy provided no evidence of a significant difference. These effects or lack of effects can also been seen in the box plots of elevation accuracy in Appendix AI. The training intervention, specifically the frame-of-reference training, was expected to improve elevation accuracy. The evidence appears to support this expectation.

The results of H10 provided no evidence that the training intervention had a significant effect on dimensional accuracy. No significant interactions, main, or simple effects were seen. Although the untrained evaluators' mean DA improved from first to second evaluation in six out of seven categories, none of these were statistically significant. The trained evaluators' mean DA improved from first to second evaluation in only three out of seven categories and again, none of these were statistically significant.

The results of the content analyses of the evaluators' first and second evaluation comments on strengths, areas for improvement, and site visit issues were summarized in Table 25. No statistical tests were performed on these data. Table 25 provides a comparison of the summary statistics before and after training. The %Hits and %Experts are presented for combined first evaluation comments, then the untrained evaluators' second evaluation comments, followed by the trained evaluators' second evaluation comments. Review of the data provided some faint patterns. Surprisingly, the %Hits and %Experts were often (more than 50% of the time) lower for the second evaluations than for the first. This may have been due to subjects tiring of the task of writing out comments. As with the first evaluation, the highest %Experts observed during the second evaluation were from the items of Category 1.0. The highest %Hits observed for the second evaluation were also from the items of Category 1.0 and all the highest statistics seen[12] were from the trained evaluators.

[12] The highest values observed were: Strengths, %Hits = 68% (Item 1.2); Strengths, %Experts = 51% (Item 1.3); Overall, %Hits = 48% (Item 1.2); and Overall %Experts = 20% (Item 1.3).

There appears to have been a degradation in evaluator performance from the first to the second evaluation; however, the training intervention seems to have lessened the degradation. As mentioned in the results of the first evaluation, averages of the %Hits and %Experts statistics must be interpreted cautiously. The average untrained subject on the typical item identified 28% of the strengths identified by the experts and 13% of the overall comments of the experts compared to 30% of the experts' strengths and 15% of the experts' overall comments for the average trained subject. 26% of the average untrained subject's comments were hits when identifying strengths and 23% were hits overall compared to 28% hits for strengths and 23% hits overall for the average trained subject. Although a degradation was seen from the first to the second evaluation, the trained subjects' averages were higher for each statistic and all the highest values observed during the second evaluation were from the trained subjects. This might be viewed as mild evidence of the effect of the training intervention.

Question 11 - How might the training of evaluators be improved?

The results of testing hypotheses 11 and 12 provided very little information for addressing Q11. Hypothesis 11 was expected to show a negative relationship between perceived difficulty in evaluating and actual accuracy. For H11, only two significant results were observed out of twenty-eight tests. One of these was for the combined subjects' first evaluation of Category 6.0 when elevation was the measure of accuracy. The other significant result[13] was for the trained subjects' second evaluation of Category 2.0 when dimensional accuracy (DA) was the measure of accuracy. For H12, only two significant results were observed out of twenty-eight tests. One of these was for the combined subjects with DA as the measure of accuracy and the other was for the trained subjects with elevation as the measure of accuracy. With a type-I error rate of 0.05, the significant results for H11 and H12 could easily have been due to chance. A likely

[13] Mild evidence of a negative relationship ($p = 0.06$, $p = 0.075$) was seen for two of the other six categories evaluated by the trained evaluators when DA was the measure of accuracy.

explanation for these non-significant results was the small sample sizes, particularly for the trained subjects. Even when using the combined data from the first evaluation, cells with as few as zero, one, or two observations were common.

The questionnaire data used to test H11 provided some useful information regarding the subjects' perceptions of which categories were most difficult[14] to evaluate. Sixteen evaluators of Category 5.0 during the first evaluation gave it a mean rating (perceived difficulty = 3.50) implying it was the most difficult category to rate. Eleven of the trained subjects from the second evaluation gave Category 5.0 a mean rating (perceived difficulty = 2.82) that was the third most difficult for their group. Category 4.0 received the most difficult mean rating (perceived difficulty = 3.14, n = 7) from the trained subjects' second evaluation and the second most difficult mean rating (perceived difficulty = 3.15, n = 20) from the combined subjects' first evaluation. Category 2.0 received the most difficult mean rating (perceived difficulty = 3.13, n = 8) from the trained subjects' second evaluation and the third most difficult mean rating (perceived difficulty = 2.95, n = 20) from the combined subjects' first evaluation. The mean perceived difficulty of the remaining categories ranged from 2.44 to 2.88, between somewhat easy and somewhat difficult.

The perception during the first evaluation that Category 5.0, Process Management, was difficult to evaluate was surprising. The fact this strong perception was not seen in the second evaluations of the trained subjects may indicate this was a random outlier; although the magnitude of the perception was by far the strongest observed. Nothing in the training intervention provides an obvious explanation of the trained subjects' perception of less difficulty in evaluating Category 5.0. The primary examples used in the training intervention were items from Categories 1.0 and 6.0. If a major change were expected for specific categories, it would be Categories 1.0 or 6.0.

[14] The evaluators rated each category they scored as: easy (1), somewhat easy (2), somewhat difficult (3), or difficult (4); therefore, a higher rating was perceived as more difficult.

The perception that Category 4.0, Human Resource Development and Management, was somewhat difficult to evaluate was not surprising. The subjects were predominantly engineers and human resources is not a typical area of expertise for engineers. Also, the content of Category 4.0 is arguably the most qualitative of the seven categories, whether the evaluator is an engineer or not. The perception that Category 2.0, Information and Analysis, was somewhat difficult to evaluate does not have such an obvious explanation.

The questionnaire data used to test H12 provided a little useful information regarding the subjects' perceptions of which categories they evaluated most accurately[15]. Both the combined subjects during the first evaluation and the trained subjects during the second evaluation indicated that they felt their scores on Category 6.0, Business Results, were somewhat close to the experts' scores. This is not surprising given that Category 6.0 is arguably the most quantitative and objective of the seven categories. . The mean perceived accuracy of the other six categories ranged from 2.32 to 2.86, between somewhat remote and somewhat close.

Question 12 - Which evaluator characteristics best predict the accuracy of the evaluators' scores?

The results of the analyses for Q12 produced equations that predict accuracy based on the evaluator's characteristics. The equations are described below, along with discussion of potential reasons for the characteristics in the equations being identified as better predictors.

The equation for the prediction of elevation accuracy explained 21.4% of the variance in elevation observed among 42 subjects. These 42 subjects were felt to be more representative of evaluators in training than the entire sample because they excluded extreme outliers and those without any quality and productivity training in the last ten

[15] For H12, the evaluators rated each category they scored in terms of how close they felt their scores were to the experts' scores: remote (1), somewhat remote (2), somewhat close (3), or close (4).

years. The 42 subjects used to produce this equation had an average of 8.1 years of work experience, 79% had completed their highest degree in engineering, and 19% had completed a master's degree. All of the subjects had completed at least a bachelor's degree. The regression equation for predicting elevation is:

$$\text{Eavg} = 21.68 - 6.46\ \text{engrdeg} + 8.68\ \text{degreeE} + 0.527\ \text{expE2}$$

Eavg represents the evaluator's average elevation on two categories the evaluator was randomly assigned. engrdeg represents whether or not the evaluator's most recent degree completed was in engineering. Thus, an evaluator whose most recent degree was in engineering is predicted to have an Eavg approximately 6.5 points lower (more accurate) than those whose most recent degree was not in engineering. Given that poor elevation was generally associated with a leniency effect, it could be that engineers are more severe "graders," thus resulting in better elevation in this case. The competitive nature of engineering school, coupled with the perceived objectivity of engineering methods are possible explanations for this relative severity. degreeE represents the highest degree completed by the evaluator, where a bachelor's degree = 0 and a master's degree = 1. Thus, evaluators that have completed a master's degree are predicted to be less accurate, in terms of elevation, than those that have not completed a master's degree. The reason for this relation is unclear. expE2 represents the evaluator's years of professional experience. Thus, evaluators with more experience are predicted to be less accurate, in terms of elevation, than those with less experience. This could be a function of more experienced evaluators being more lenient. Observations from the correlation analysis implied that older, more experienced evaluators are likely to be more lenient (resulting in poor elevation accuracy), but better able to discriminate between relative strengths and weaknesses (resulting in better dimensional accuracy).

The equation for the prediction of dimensional accuracy explained 53.8% of the variation in dimensional accuracy observed among 53 subjects. These 53 subjects had an average of 6.9 years of work experience and an average of 1.2 years working in QC/QA/QI. 26.4% of the subjects described themselves as full time students. The 53 subjects described their employers as follows: manufacturing (15%), service (24.5%), federal government (32%), and state or local government (28%). 41.5% of the subjects were from large (over 5000 employees) organizations, 39.6% from medium size (between 501 and 5000 employees) organizations, and the remainder, 18.9%, were from small organizations (1 to 500 employees). Like before, these subjects were felt to be more representative of evaluators in training than the entire sample. Outliers and those who reported their employer's number of employees as zero (i.e., the unemployed) were dropped. The regression equation for predicting dimensional accuracy is:

$$\begin{aligned} \text{DAavg} = 25.3 - 0.525\ \text{exp3D} - 6.29\ \text{std} - 2.80\ \text{mfg} - 5.23\ \text{svc} \\ + 0.631\ \text{yrsqcqa} - 3.73\ \text{size3D} \end{aligned}$$

DAavg represents the evaluator's average DA on two categories the evaluator was randomly assigned. exp3D represents the evaluator's years of professional experience. Thus, evaluators with more experience are predicted to be more accurate, in terms of DA, than those with less experience. An experienced evaluator may be more familiar with the details and subtleties of the items being evaluated[16] and thus better able to distinguish between relative strengths and weaknesses. std represents whether or not the evaluator is a full-time student. It was surprising to see full-time students predicted to be over 6 points more accurate, in terms of DA, than those who are not full-time students. Perhaps full-time students are better read and more familiar with the subtleties of the items being evaluated. Although full-time students were positively correlated with having completed a master's degree (i.e., they are more likely to be Ph.D. students), this was not associated with better DA. mfg (manufacturing) and svc (service) were indicator variables describing the evaluator's employer. All other subjects described their employers as either fed (federal) or stloc (state or local) government. Earlier versions of the equation included fed and stloc with positive coefficients, but excluded mfg and svc. Thus, evaluators working for manufacturing and service organizations were predicted to have better accuracy, in

[16] An assumption for the remainder of this discussion is that those more familiar with the details and subtleties of the items being evaluated are more likely to be able to distinguish between the relative strengths and weaknesses displayed on these items (i.e., they will be more accurate in terms of DA).

terms of DA, than those working for government organizations. Although government organizations, particularly federal government organizations, have long been involved in quality and productivity improvement, the extent of implementation and deployment appears more advanced in the private sector. This might be attributed to competitive pressure (i.e., do or die) in the private sector. This forcing function may have required those employed in the private sector to be more familiar with the content of the items being evaluated. yrsqcqa represents the number of years the evaluator has worked in the quality control, quality assurance, or quality improvement function. It was surprising to see those with more experience in QC/QA/QI predicted to be less accurate, in terms of DA, than those with less experience in QC/QA/QI. The traditional view of quality control and quality assurance (Garvin, 1988) is narrower than that implied by the criteria of the Baldrige Award. It may be that inconsistencies in these perspectives caused those with more experience in QC/QA/QI to be less accurate in evaluating an organization against categories of the Baldrige Award. The final variable in the equation, size3D, represents the number of employees in the evaluator's employing organization coded from 1 (a small organization) to 4 (a large organization[17]). Thus, evaluators working for larger organizations were predicted to be more accurate, in terms of DA, than those working for smaller organizations. Interestingly, most (about 75%) of those working for larger organizations worked for the federal government[18]. Thus, among evaluators working for large organizations those working for the private sector are predicted to be more accurate, in terms of DA. Among evaluators working for the public sector, those working for larger organizations (i.e., federal or state government) are predicted to be more accurate. On the other hand, an evaluator from a large government organization would be predicted to be more accurate than an evaluator from a small private organization. Thus, the size of the employing organization appears to be a more important predictor than the type of

[17] The variable size3D dropped those reporting their employer's number of employees as zero.

[18] Subjects who worked for federal or state government were not asked to give a specific number of employees for their employer. Instead, they were automatically placed in the largest employing organization size category.

organization. Could it be that larger organizations are more likely to have adopted a formal quality and productivity improvement effort that would increase the evaluator's familiarity with the content of the items being evaluated?

Table 33 provides a concise summary of the results of the analyses. Table 33 corresponds to Table 1: The Structure of the Research Problem. Whereas Table 1 listed the data needs and analysis procedures for each research question and research hypothesis, Table 33 lists a summary of the results from each analysis.

Table 33. Summary of the Results

Research Questions	Hypotheses	Summary Results
For a given sample of untrained evaluators, what is the consistency of their scores?	H1: There will be no difference in scores between the treatment and control groups during the initial evaluation: - a) by item, and - b) by category.	No statistical differences in mean scores by item or category. Practical differences ranged from 0.0 (Item 4.1) to 13.3 (Item 2.3) for items and 2.1 (Category 6.0) to 7.6 (Category 2.0) for categories.
Q1: How much agreement is there among evaluators on the score of an item?	H2: For each item, there will be no difference in score variances between the treatment and control groups.	No statistical differences in item score variances. Range of std. deviations = 8.4 to 26.3 for T1 and 9.4 to 25.2 for C1.
Q2: How much agreement is there among evaluators on the score of a category?	H3: For each category, there will be no difference in score variances between the treatment and control groups.	No statistical differences in category score variances. Range of std. deviations = 9.2 to 17.4 for T1 and 7.7 to 18.4 for C1.
Q3: How consistent is the within-item variation of evaluator scores across all the items of a category?	- na	Boxplots of item scores show fairly consistent variation for Categories 1.0 and 3.0, somewhat consistent variation for Categories 5.0 and 6.0. Different constructs within Category 7.0 were highlighted by the pattern of score dispersion. Categories 6.0 and 7.0 appear to have smaller score variances than the other categories. See Figure 5 and Table 7.
Q4: How consistent is the within-category variation of evaluator scores across all seven categories?	H4: There will be a difference in score variances between categories.	
For a given sample of untrained evaluators, what is the accuracy of their scores? Q5: How accurate are the evaluators' scores for each category?	H5: There will be no difference in accuracy between the treatment and control groups during their first evaluations:	Mean elevation ranged from 18.8 (c4.0) to 39.0 (c3.0) for the control and 16.1 (c7.0) to 33.1 (c7.0) for the treatment group. Mean DA ranged from 7.8 (c2.0) to 18.1 (c5.0) for C1 and 7.5 (c6.0) to 14.1 (c5.0) for T1. No statistical differences in mean elevation or mean DA for any category. Boxplots (Fig. 6) indicate a leniency effect. Content Analysis: %Experts ranged from an avg. of 30% on Item 1.1 to 6% on Item 7.5. The average subject on the typical item identified 36% of the strengths identified by the experts and 17% of the overall comments of the experts. 35% of this average subject's comments were hits when identifying strengths and 28% were hits overall.

table continues

Research Questions	Hypotheses	Summary Results
For a given sample of untrained evaluators, will evaluator training change the consistency of their scores?	H6: There will be a difference in scores between the treatment and control groups during the second evaluation: - a) by item, and - b) by category.	Leniency effect reduced for trained evaluators. Most Items of Categories 5.0 and 6.0 show evidence of an interaction. Items 5.3, 5.4, 6.1, 6.2, and 6.3 displayed significant simple group effects after training (none pre-). Items 6.1-6.3 displayed significant simple time effects for the treatment group.
Q6: Did agreement among evaluators on the score of an item change (improve) due to evaluator training?	H7: Item score variances will be smaller for second evaluation scores than for first evaluation scores.	No differences for treatment group. Three items showed significant differences for the control group, believed random. No differences between groups after training.
Q7: Did agreement among evaluators on the score of a category change (improve) due to evaluator training?	H8: Category score variances will be smaller for second evaluation scores than for first evaluation scores.	Range restriction near upper bound caused control group variances to decr. 2 out of 3 times their central tendency increased. These 2 were the only significant differences observed.
Q8: Did within-item variation of evaluator scores across all the items of a category change (decrease) due to evaluator training?	- na	No patterns observed. See Figure 8 boxplots.
Q9: Did within-category variation of evaluator scores across all seven categories change (decrease) due to evaluator training?	-H9: There will be a difference in score variances between categories for both the treatment and control groups.	Training appears to have reduced differences across categories below statistical significance (possible sample size issue). A significant difference (p<0.01) was seen across categories for the control group, reinforcing the results of H4.
For a given sample of untrained evaluators, how will evaluator training change the accuracy of their scores?		Training appears to improve elevation accuracy. Only one interaction (c5.0) was seen, but significant simple group effects were seen in 5 of 7 categories after training. Significant simple time effects seen in 2 categories for the treatment group (none for control).
Q10: Did the accuracy of the evaluators' scores change (improve) due to the evaluator training?	H10: The accuracy of evaluators' scores will improve between the first and second evaluations.	No evidence of effects on DA.

table continues

Research Questions	Hypotheses	Summary Results
Q11: How might the training of eval·ators be improved?	H11: Subjects' perceived difficulty of evaluating a category will be negatively related to their accuracy in evaluating that category.	No evidence of a relation. Possible sample size (observations per cell) problem. Category 4.0 perceived as relatively difficult to evaluate (most difficult by T2 and 2nd most difficult by CT1).
	H12: Subjects' perceived accuracy in scoring a category will be positively related to their accuracy in evaluating that category.	No evidence of a relation. Again, sample size resulted in small numbers of observations per cell in the ordered categories matrices.
Q12: Which (of the following) evaluator characteristics best predict the accuracy of the evaluators' scores? - previous use of MB criteria - level of education - educational specialty - amount of Q&P training - work experience - job function - job level - employer industry - employer size - age - gender	- na	Using a subset of 42 out of 67 respondents, a prediction equation for elevation was developed with an adjusted R-square of 0.214. This model needs is of questionable utility. Using a subset of 53 out of 67 respondents, a prediction equation for DA was developed with an adjusted R-square of 0.538. This subset dropped outliers and those reporting their employer size as zero (i.e., unemployed). Each of the variables in the equation, except QA experience, has a negative relationship with the DA index (i.e., represents improved accuracy). DAavg = 25.3 - 0.525 exp3D - 6.29 std - 2.80 mfg - 5.23 svc + 0.631 yrsqcqa - 3.73 size3D

VI. Conclusions

This chapter presents the conclusions of this research. Chapter IV contained the analysis of the data. Chapter V presented the results of the analyses and some interpretation of these results. This chapter goes one step further, presenting an integration of information across questions and hypotheses. The chapter begins with overall conclusions. The next section discusses specific research conclusions, followed by the researcher's observations and opinions. The chapter ends with a brief summary.

Overall Conclusions

This study provided several contributions to the body of knowledge regarding third-party evaluators scoring of written organizational self-assessments. A public baseline or benchmark was established that can be used to compare the scores of other groups of evaluators. Even in cases where other scales (i.e., scoring guidelines) are used, the relative magnitude of the consistency of their scores might be evaluated against this baseline. The use of accuracy indices, borrowed from performance appraisal research, was demonstrated as a way to assess and compare the evaluators' scores. Operational examples were developed to aid in the interpretation of the meaning of a particular accuracy index value. These accuracy indices, the baseline of index values created, and the operational examples offer new tools to those who train and assess evaluators. The effects of a moderate (2.5 hours) FOR training intervention on the scores of the evaluators were demonstrated. The family of statistics calculated from the scores before and after the training intervention provided a detailed illustration of these effects. Examination of the data found a number of implications for improving evaluator training[1]. Some of these findings have implications for those who teach content similar to that taught in this training intervention (i.e., the Baldrige criteria or quality management). Finally, this study

[1] These implications are particularly relevant for training evaluators (examiners) for the Baldrige Award or similar quality/productivity awards.

produced two empirical models of which evaluator characteristics best predict the accuracy of evaluators' scores. While these models are preliminary, they are based on quantitative data and provide a starting point for future research.

Research Conclusions

The pre-training analyses showed wide fluctuations in within-group dispersion and between group location and dispersion of the evaluators' scores; however, no statistical differences were observed. In hindsight, all the statistical tests performed on the pre-training scores may not have been necessary. On the other hand, these analyses demonstrated there were no statistical differences between the groups even though the location and dispersion of their scores often appeared to differ widely. One interesting pattern emerged from the review and testing of the pre-training data. Comparison of within-category score variation showed categories with more quantitative content (Categories 6.0 and 7.0) had significantly less score variance than those with more qualitative content[2] (the remaining categories). This has implications for training evaluators and perhaps the education and training of those studying content similar to that used for this experiment (i.e., quality/productivity management). The content areas (categories) with more score variance (i.e., the more qualitative) might require more time and emphasis if the objective is to bring all students (evaluators) to some common level of understanding.

The accuracy indices appear to be useful for assessing the rating effectiveness of evaluators. Elevation is useful when the scores are being used to see if the organization meets some minimum threshold or level of performance. In these situations it is important that the evaluator's overall score, usually an average or weighted average of dimension scores, be close to that of the true score. While each dimension score might be off by plus

[2] This pattern was not observed in the post-training scores of the trained and untrained evaluators (the groups' pre-training scores were combined for this analysis). This may have been due to the smaller sample sizes (average n = 10 versus 22), the effects of range restriction on the untrained evaluators' scores, or the effects of the training on the trained evaluators' scores.

or minus a few percentage points, on average they should be close to the true scores. Elevation provides an indication of how close an evaluator's overall score is to the true score and might be useful for determining when evaluators' scores are accurate enough to be used for decision making. Dimensional accuracy is useful when scores are being used to provide the organization with feedback on relative strengths and weaknesses. In these situations it is important that the evaluator's scores on each dimension reflect the relative strength or weakness of the organization on that dimension. Thus, the organization should be able to target the dimensions with relatively low scores as areas needing improvement and be confident that these are the areas with the most potential for improvement. Dimensional accuracy provides an indication of how well the evaluator can discriminate between the relative strengths and weaknesses of a set of performance dimensions and might be useful for determining when evaluators' scores are accurate enough to be used for detailed feedback. In terms of Conti's (1994) two purposes of self-assessment, elevation appears more important for conformity self-assessment and dimensional accuracy appears more important for improvement-oriented self-assessment.

A moderate frame-of-reference (FOR) training intervention does appear to effect the scores of evaluators, but not necessarily in the ways expected. FOR training appears to improve elevation accuracy, but appears to do nothing for dimensional accuracy. Given the leniency effect observed for the untrained evaluators, it is not surprising that training aimed at improving the evaluators' frame-of-reference would improve elevation accuracy[3]. On average, the training intervention brought the evaluators' scores more in line with the scores of the experts (i.e., the true scores); however, the evaluators' ability to discriminate between the relative strengths and weaknesses of related dimensions (items) did not improve. Performance appraisal studies have found FOR training to improve dimensional accuracy (Stamoulis and Hauenstein, 1993; Pulakos, 1986). Given the number and complexity of the dimensions used in this experiment, it is likely that more training on the

[3] Without an obvious leniency (or severity) effect in the pre-training scores, the effect of the FOR training on elevation might have been less pronounced. In such a case, a larger sample size might be required to demonstrate the statistical significance of the change in elevation.

dimensions themselves[4] (PDimT) is required than that required for a similar performance appraisal experiment (e.g., evaluation of the performance of a classroom instructor or an interviewer).

FOR training appears to have the biggest impact on the dimensions with more quantitative content. When differences in mean scores were tested by H6, most of the significant results were clustered around Categories 5.0, 6.0 and 7.0. Category 6.0, arguably the most quantitative of the seven categories[5], clearly showed the most significant changes. Like the difference in score variances between categories, this has implications for the training of evaluators and perhaps the education and training of those studying content similar to that used for this experiment. The FOR training seems to have the greatest effect on the mean scores of the more quantitative categories and these were the categories with the least variation in their scores before training. Like before, this implies a need to emphasize the content of the more qualitative dimensions (categories) when training evaluators. This observed effect may in part be explained by the subjects used for this experiment. Three-quarters of the subjects said the last degree they completed was in engineering. Future research might seek a more diverse subject pool to see if the effects are still more pronounced for the relatively quantitative dimensions.

FOR training does not appear to improve agreement among evaluators on the score of a particular dimension (e.g., item or category). While the training intervention resulted in a shift in mean scores (toward the true scores), it did not appear to reduce the variance of the scores. This might be explained by the range restriction caused by the leniency effect observed for the untrained evaluators. Thus, the variance of the untrained evaluators' scores tended to be constrained due to proximity to the upper bound. The variance of the trained evaluators' scores was less constrained since their scores tended

[4] An assumption underlying the choice of the two courses from which subjects were drawn was that these courses (and the associated curricula) included content related to most of the Baldrige criteria. While this assumption may be correct, more training on the meaning and applicability of each item in the Baldrige criteria may be necessary to significantly improve dimensional accuracy.

[5] It is interesting to note that the 1995 Baldrige criteria weight Category 6.0 (and Category 7.0) with the most points (250 each out of 1000 total).

toward mid-range. Future research might use a case where the average true scores are in the lower quartile of the range of scores. This would result in a tendency to constrain scores moving toward the true scores (i.e., those of the trained evaluators) and have little effect on scores that are changing randomly from the first to second evaluation (i.e., those of untrained evaluators). The results of such an experiment could be compared to the results of this experiment to see if the effects of training and range restriction near the true scores result in decreasing variance for more items and categories. Another possibility for future research is to use a second round of scoring where small groups (e.g., 3 evaluators) produce a group score for the items they scored individually. Previous research on group decision making would likely suggest this second round of scoring will reduce variance, but what effect might it have on accuracy? Will group scoring affect elevation and DA differently?

FOR training appears to reduce within-category variation of evaluator scores relative to the variation of the other categories. That is, within-category score variation appears to be more consistent across the seven categories as a result of the training intervention. While the pre- to post-training comparison of category score variation did not show a statistically significant difference for any specific category, the relative difference in category score variation across the seven categories did appear to decrease as a result of the training. A difference in the score variation from category to category might be expected due to differences in category content. This expected difference was seen and was statistically significant for the combined first evaluation category scores and the untrained evaluators' second evaluation category scores. The differences in category score variance for the trained evaluators' second evaluation (post-training) were not statistically significant. This provides evidence of a mild effect of FOR training on the variance of evaluator scores. A stronger effect might have overcome the problems of the leniency effect coupled with range restriction described above and resulted in significant differences in comparisons of pre- and post-training score variances for each category.

The accuracy prediction equations developed in response to Q12 may be viewed as preliminary models for selecting evaluators[6]. These models should be tested and refined with data from other samples of evaluators. A potential follow-up to this study is to see how well these equations predict the evaluators' accuracy on the second evaluations. The equations should work relatively well for predicting the untrained evaluators' accuracy from their second evaluations. The equations may not work as well for predicting the trained evaluators' accuracy from the second evaluations, particularly for elevation. Since most of the effects observed related to elevation, the equation's predictive effectiveness would likely change. Also, the equation for predicting average elevation was the weaker predictor of the two equations (R-squared adjusted of 0.214 versus 0.538). The equation for predicting average elevation is of limited value at this point. It only explains 21% of the variation in elevation and it contains variables that may be sample dependent. The equation for predicting average DA is more promising and provides a basis for a model of evaluator characteristics that affect the accuracy of their scores. It explains over 53% of the variation in DA and most of the variables in the equation are generalizable to more diverse samples. Intuitively, most of the variables and coefficients in the DA equation make sense (see the discussion of results in Chapter V).

A number of lessons were learned in the execution of this experiment. These were documented and shared with the instructors who collaborated in this experiment shortly after the completion of data collection (see Table 34). These lessons have implications for future research and for the training of evaluators. The increased integration of the evaluation into the course (second bullet in Table 34) is consistent with providing more performance dimensions training (PDimT) as part of the treatment. Given the baseline established by this study, perhaps a non-experimental design could be used where all subjects receive the treatment. The results could be compared to this study to see if the magnitude of the effects was increased by the more in-depth training intervention.

[6] It may be more appropriate to say these are preliminary models for selecting evaluators in training.

Table 34.

Lessons Learned from Conducting the Experiment (12/20/95)

- I would trade off generalizability for potentially improved responses by providing 15-20 minutes of detailed instructions along with the first Scorebooks. Although basic instructions were given orally and specific instructions were given in writing, further explanation was scattered throughout the criteria booklets and Scorebooks. Oral description of how the tasks interrelate and what responses might look like could be provided without jeopardizing the impact of the later frame-of-reference training (i.e., treatment). While this might limit generalizability to the truly self-taught, it would be comparable to an organizational approach where a central coordinator studied the evaluation process and provided an overview to the evaluators.

- I would increase integration of the experiment/evaluation into the course. This would require all materials be developed several weeks (or months) prior to the beginning of the course. I would include some usage or assigned reading from the Baldrige Criteria booklet early in the course. The tasks of the evaluation might be further spread out or broken into smaller pieces. Perhaps the sequencing of the evaluation of selected categories could coincide with similar topics discussed in class. Spreading the evaluation over a longer time period might allow each student to evaluate the entire case study (i.e., all seven categories). A group or team portion of the evaluation might also be incorporated. The weighting of the evaluation as part of the course grade should be better balanced relative to the other assignments in the course. In the ideal situation, the researcher might also be the instructor.

- I would avoid using a televised course. While the actual lectures seemed to go well, controlling distribution and collection of materials on time was often difficult. The interaction during the class evaluation of the sample case excerpts seemed to work well from both the instructor and students' perspective. Controlling the distribution of materials via numerous site coordinators was less than ideal. Sending materials well in advance at least once resulted in premature distribution. Timely collection is difficult due to the varied schedules of the off-campus students and the wide variation in submit to receipt time.

table continues

- I would have made assignment to the control or experiment group more explicit. In an effort to reduce bias, I purposely did not use these terms or explain the different assignments in advance. The subjects were aware this was part of a research project (and were quite supportive). They were also told that some others in the class were given a different task (schedule) during the second evaluations; however, the relation of these differing tasks to the research design was not explained. They were not informed of this difference until the second evaluations were actually distributed. With the additional instruction described in the first bullet above, I believe the dysfunctional consequences of letting subjects know they are being randomly assigned to a group could be limited. I would not tell subjects which group they were assigned to, simply that this assignment would occur. Subjects could then figure out which group they were in once the second evaluations began.

Suggestions for future research have been sprinkled throughout the conclusions to this point. Other suggestions include: (1) A comparison of FOR training to PDimT to identify differential effects and the implications for how best to spend limited training time; (2) An experiment where all the subjects evaluate the same categories. This could be done by having one-half the subjects evaluate category A and the other half evaluate category B during the first evaluation. Following the treatment, those who had already evaluated category A would now evaluate category B and vice versa. This would increase sample size and eliminate inherent categorical differences; (3) Comparisons of the data from samples of experienced evaluators to the results of this study. If a moderate training intervention can significantly improve the scores of previously untrained evaluators, how much better are the scores of experienced[7] evaluators? and (4) Examine the underlying measurement systems of discriminating[8] evaluations such as supplier certification, accreditation, and ISO 9000 certification to see if accuracy indices can be calculated and studied. Since these evaluations determine whether or not an organization meets some minimum level of performance, issues of score consistency and elevation accuracy are worthy of further investigation.

Researcher's Observations

This section contains the researcher's observations that may not be reflected in the previous conclusions. These observations were not necessarily supported by the data, but represent the researcher's educated opinions after having conducted this study.

The qualitative dimensions need more emphasis than the quantitative dimensions when conducting content training (PDimT) for evaluators. While such training may reduce the variance in the scores of these qualitative dimensions, they are likely to still

[7] The expert scores used for this study were from a team of Baldrige Award Senior Examiners, representing some of the best trained and most experienced examiners available. The suggestion here is to compare the results of this study to the scores of more typical evaluators, such as examiners for a Corporate or State Quality Award or even new Baldrige Examiners.

[8] Discriminating evaluations are similar to Conti's conformity self-assessments; however, his definition conformity self-assessment was deemed to narrow (see Chapter 2).

display more variance than the quantitative dimensions (see next paragraph for further explanation).

The smaller score variances observed for the more quantitative categories of the Baldrige Award criteria might be viewed as support for weighting these categories more heavily than the others. In the 1995 Baldrige Criteria, Categories 6.0 and 7.0 represent 50% of the overall score. The content of Categories 6.0 and 7.0 are focused on end results, including profitability, market share, and customer satisfaction. The desirability of these results is widely understood. This common understanding partially explains why the scores of these categories are more consistent than the other categories. The cause and effect relationships between activities (i.e., causes) falling under the more qualitative categories and end results (i.e., effects) is not always clear or widely understood. Thus, the desirability of these activities is not as widely understood as those that relate directly to profitability or customer satisfaction. This may explain why the scores of the more qualitative categories are less consistent. This also justifies weighting the scores of the qualitative categories less than the scores of the quantitative categories. Because we have less faith in the scores from these qualitative categories, we would not want their scores to overly influence our final decision.

From a training perspective, poor elevation may be easier to correct than poor DA when the poor elevation is related to a common leniency or severity effect. That is, feedback after training and assessing the evaluators' performance could let them know that their scores tend to be x points high or low[9]. In this study, the initial elevation accuracy was very poor[10] (average elevation approximately 24) while the DA was mediocre (average DA approximately 11). The illustrations in Figure 15 show what this level of accuracy looks like. So while poor elevation may be easier to correct, poor elevation was also the bigger problem.

[9] This would be similar to rater error (RE) training, except RE training is commonly given without the benefit of knowing whether a particular evaluator suffers from a leniency or severity effect. In RE training, these effects are explained and the evaluator is admonished to avoid them. The proposal here is to conduct RE training after a thorough analysis of the evaluators' post-FOR & PDimT scores.

[10] Plotting the data showed the poor elevation to be related to a leniency effect.

Improvements in either index are likely to become more difficult as the average accuracy improves (i.e., approaches zero). The average Elevation after training was better (about 17), but still needs improvement from a practical perspective. The average DA after training remained at around 11[11]. A more involved training intervention is needed to improve DA. That is, from the trainer's perspective, DA may be the more challenging aspect of accuracy to improve. From the ratee's perspective, elevation may be more important, particularly if there are consequences based on the ratee's overall mean score.

This research has potential application with a variety of self-assessments that are scored by third-party evaluators. For quality and productivity awards like the Baldrige and internal company awards, the procedures developed for this research could be used to diagnose and treat rater errors. Assessment of the accuracy of evaluators' scores could be used to determine when an evaluator is certified to be accurate enough to evaluate organizations. For example, knowing which evaluators' scores tend to have relatively poor DA may be useful for selecting who is to give feedback to the ratee organization. While an evaluator with poor DA might be a contributing member of a team of evaluators and might "know a good organization when they see it" (i.e., have good elevation accuracy), they might not be able to give detailed feedback on which dimensions are relative strengths and which are relative weaknesses.

In addition to quality and productivity awards, this research is applicable to ISO 9000 and ISO 14000 certification (and the training of the registrars), academic and other forms of accreditation, and supplier certification. According to ASQC, there is no official body for accrediting ISO 9000 registrars in the United States. Some registrars have obtained accreditation from European countries and use this as a surrogate for U.S. accreditation. Others obtain accreditation from the private Registrar Accreditation Board (RAB), but this may or may not become the official accrediting body. In the meantime,

[11] A DA index value of 11 is not ideal, but may be useful for giving feedback. If the subject's item scores in Figure 15 are reduced by 23 points to account for the leniency effect, the resulting item scores provide some indication of which items are relative strengths and which are relative weaknesses. Only item 7.4 or one out of five would be problematic (i.e. too inaccurate for useful feedback).

registrars might assess their evaluators using methods like that used in this research. Summaries of these assessments could be used to demonstrate the competency of their evaluators. The ISO 14000 environmental management standards are still being developed, but their certification process might also benefit from this approach.

Academic accreditation may need the findings of this research even more than other forms of assessment. Academic evaluators (visitors) often evaluate a program in small teams of only one or two visitors[12]. When a program is being evaluated by a single visitor, the consistency of scoring among visitors is critical. A wide variance among visitors could result in the outcome of the accreditation depending upon which visitor is assigned to a particular program. Academic accreditation appears to depend on the visitors' paradigms of what an academic program should look like (i.e., very much an institutional view of the academic organization).

ABET (the engineering accreditation body) is revising their criteria and assessment procedures to be more like the Baldrige Award. Training of ABET visitors appears to vary from discipline to discipline. For Industrial Engineering visitors, the training is brief[13] and seems to depend heavily on the visitors' prior knowledge and experience. Typically, ABET visitors are expected to participate in the accreditation of an engineering program as observers for part of their training. This is a classic example of what Deming (1986) referred to as "worker training worker." Evaluator training that depends on worker training worker and the evaluator's prior knowledge, experience, and paradigms is not likely sufficient to move ABET to a successful Baldrige type assessment.

Group consensus scoring is likely to improve DA, but may have little effect on elevation when a leniency or severity effect is widespread. As shown in Figure 14, averaging the control group's scores on the items of Category 7.0 resulted in relatively good dimensional accuracy (DA = 3.8), but poor elevation (E = 22.2). The leniency effect

[12] These visitors may be part of a larger team that is evaluating multiple programs across the institution; however, each program is evaluated somewhat independently.

[13] The researcher attended this training in 1995 and found it lacking. It did not include any form of frame-of-reference training, performance dimensions training, or rater error training.

appears to affect most of the control group; therefore, averaging their scores does not reduce the leniency nor improve elevation. On the other hand, their average scores do appear to discriminate between the relative strengths and weaknesses of the items and thus, result in good DA. A follow up study that includes group consensus scoring could be used to see if DA improves to desirable levels. Coupled with FOR training, PDimT, and post-training (RET) feedback, group consensus scoring could result in third evaluation[14] scores with greatly improved elevation and DA.

The exploratory regression analyses produced some interesting intermediate results (see Appendices AY and AZ). The evaluators with a lot of work experience (e.g., over 15 years) appeared to have relatively poor elevation accuracy, but relatively good dimensional accuracy compared to the other evaluators. That is, they tended to be too lenient in scoring, but they discriminated well between relative performance on related items. A similar trend was observed for the supervisors with more supervisory responsibility when compared to the other supervisors. This could be a reflection of the wisdom gained from substantial experience in working with people. This might also be a reflection of extensive experience with performance appraisal systems. That is, giving everyone high scores typically results in the same consequences for everyone[15] and reduces the chance of unnecessarily offending anyone. Having neutralized the potentially dysfunctional consequences of a low (or even moderate) score, the ratee is more likely to be acceptant regarding feedback on relative strengths and weaknesses. Within the microcosm of the immediate workplace, it is more important to have the ratee recognize their relative strengths and weaknesses than it is to assign a meaningful overall rating to the ratee's performance that has little impact on the consequences of the rating.

[14] This assumes pre- and post- training evaluations (like in this study), followed by rater error training and a third evaluation.

[15] Performance appraisal often requires a noticeably large difference in scores (i.e., ratings) to result in different consequences. In these cases, it may not be worthwhile to discriminate between close performing ratees since the only consequence may be to offend those who are told they scored a few points lower. In a system where a small difference in scores may result in different consequences (e.g., a forced ranking system), the evaluator may be forced to make decisions their data doesn't support. In these cases, the evaluators may learn to avoid discriminating between performers except when necessary.

Summary

This study set out to examine the impact of third-party evaluator training and characteristics on the scoring of written organizational self-assessments. But the study did more than that. It created a database illustrating the consistency and accuracy of untrained evaluators' scores where none existed before. The study showed how a common form of evaluator training, frame-of-reference (FOR) training, specifically affected the consistency and accuracy of the evaluators' scores. The training improved elevation accuracy and reduced a leniency effect, but did little to improve the agreement among evaluators (i.e., reduce the between evaluator variance) or improve dimensional accuracy. The study illustrated how inherent differences in the content of the dimensions (in this case, Baldrige Award categories) affected the dispersion of the scores. Finally, regression procedures were used to develop prediction equations for both elevation and dimensional accuracy. Although the elevation equation is rather limited, the dimensional accuracy equation appears promising for further testing and development.

VII. References and Bibliography

Accreditation Board for Engineering and Technology. (1994). Criteria for accrediting programs in engineering in the United States, effective for evaluations during the 1994-95 accreditation cycle. Baltimore.

Accreditation Board for Engineering and Technology. (1995). Criteria for accrediting programs in Engineering in the United States, EAC criteria committee draft document, March 19. Baltimore.

Aldridge, M. D. and E. S. Pape. (1994). Report of the ABET/NSF accreditation process reform workshop, August 13-14, 1994. Unpublished manuscript. Baltimore.

Anttila, J. (1994). A TQM self-assessment process using quality award criteria and a knowledge-based methodology: Case of Telecom of Finland. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 47-56). Torino, Italy: European Organization for Quality.

Aquino, M. A. (1990). Improvement vs. compliance: A new look at auditing. Quality Progress, 23(10), pp. 47-49.

Bell, D. A. and G. Wilson. (1994). The applicability of self-assessment in the context of the small organisation. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 131-140). Torino, Italy: European Organization for Quality.

Bemowski, K. (1992). Donning a new hat. Quality Progress, 25(7), pp. 21- 25.

Bemowski, K. and B. Stratton. (1995). How do people use the Baldrige Award criteria? Quality Progress, 28(5), pp. 43-47.

Bender, L. (1983). Accreditation: Misuses and misconceptions. In K. E. Young, C. M. Chambers, H. R. Kells, and Associates. Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality. (pp. 71-86). San Francisco: Jossey-Bass.

Benedetti, G. and G. Bertorelli. (1994). The IBM market driven quality assessment. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 99-108). Torino, Italy: European Organization for Quality.

Bernardin, H. J. and C. S. Walter. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62(1), 64-69.

Bernardin, H. J. and E. C. Pence. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65(1), 60-66.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of perfo⌐mance evaluation ratings. Journal of Applied Psychology, 60(5), 556-560.

Bowles, J. (1992). Debate: Does the Baldrige Award really work? Harvard Business Review, 70, (1), January-February, p. 127.

Business Week. (1991). Special Quality Issue.

Charnoff, G. (1990, December). Techniques of self assessment. Paper presented at the Executive Conference on Self Assessment for Self Assurance, American Nuclear Society, San Diego, CA.

Coakley, C. W. (1995). Class notes for STAT 5404/5644. Blacksburg, VA: VPI&SU University Printing Services.

Conti, T. (1994). Time of a critical review on quality self-assessment. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 169-180). Torino, Italy: European Organization for Quality.

Craft, A. (1992). (Ed.), Quality assurance in higher education: Proceedings of an international conference Hong Kong, 1991. London: The Falmer Press.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 52(3), 177-193.

Crosby, P. B. (1979). Quality is free: The art of making quality certain. New York: McGraw-Hill.

Crosby, P. B. (1992). Debate: Does the Baldrige Award really work? Harvard Business Review, 70, (1), January-February, pp. 127-128.

Dagnino, B. V. and J. F. B. de Souza. (1994). The use of the national quality award criteria for self-assessment in Brazil. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 215-223). Torino, Italy: European Organization for Quality.

Deming, W. E. (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology Center for Advanced Engineering Study.

Deming, W. E. (1992). Debate: Does the Baldrige Award really work? Harvard Business Review, 70, (1), January-February, p. 133.

Deming, W. E. (1993). The new economics for industry, government, education. Cambridge, MA: Massachusetts Institute of Technology Center for Advanced Engineering Study.

Dillman, D. A. (1978). Mail and telephone surveys: The total design method. New York: John Wiley & Sons.

Easton, G. S. (1993). The 1993 state of U.S. total quality management: A Baldrige examiner's perspective. California Management Review, Spring, pp. 32-54.

Eatɔn, B. D. J. (1995). Cessna's approach to internal quality audits. IIE Solutions, 27(6), 12-16.

El-Khawas, E. (1983). Accreditation: Self-regulation. In K. E. Young, C. M. Chambers, H. R. Kells, and Associates. Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality. (pp. 54-70). San Francisco: Jossey-Bass.

Evans, R. (1994). Making quality a way of life at British Airways. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 7-14). Torino, Italy: European Organization for Quality.

Federal Quality Institute. (1994). The president's quality award program 1995 application. Washington, DC: U.S. Government Printing Office.

Fuchs, E. and S. vH. Stuntebeck. (1994). The use of Baldrige-based self-assessment in AT&T. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 15-26). Torino, Italy: European Organization for Quality.

Gallagher, W. M. (1994). Self-assessment using the European Quality Award model - lessons learned by assessors. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 91-98). Torino, Italy: European Organization for Quality.

Garvin, D. A. (1988). Managing quality. New York: The Free Press.

Garvin, D. A. (1991). How the Baldrige Award really works. Harvard Business Review, November-December, 80-93.

Gibson, M. J. W., and H. Sluis. (1994). The European Quality Award process at KLM. Two years of lessons learned and results achieved. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 117-130). Torino, Italy: European Organization for Quality.

Godfrey, A. B. and D. H. Myers. (1994). Self-assessment using the Malcolm Baldrige National Quality Award. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 67-78). Torino, Italy: European Organization for Quality.

Harvard Business Review. (1992). Debate: Does the Baldrige Award really work? Vol. 70, (1), January-February, pp. 126-147.

Hays, W. L. (1988). Statistics, Fourth Edition. Fort Worth, Texas: Holt, Rinehart, and Winston.

Hertz, H. (1995). Malcolm Baldrige National Quality Award Program Office, National Institute of Standards and Technology, Gaithersburg, MD. Personal communication to G. Coleman, September X.

Jernberg, B., J. Lindström, and R. Chocron. (1994). The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 35-46). Torino, Italy: European Organization for Quality.

Juran, J. M. (1986). The quality trilogy. Quality Progress, 19(8), pp. 19-24. f

Kaplan, R. S. and D. P. Norton. (1992). The balanced scorecard - measures that drive performance. Harvard Business Review, January-February, pp. 71-79.

Kells, H. R. (1992a). Purposes and means in higher education evaluation. Higher Education Management, 4(1), 91-102.

Kells, H. R. (1992b). An analysis of the nature and recent development of performance indicators in higher education. Higher Education Management, 4(2), 131-138.

Mack, G. A. and J. H. Skillings. (1980). A Friedman-Type Rank Test for Main Effects in a Two-Factor ANOVA. Journal of the American Statistical Association, 75(372), 947-951.

Martellani, L. (1994). Self-assessment as a way for the adaptive organization. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 109-116). Torino, Italy: European Organization for Quality.

McCormick, E. J. and D. Ilgen. (1985). Industrial and organizational psychology, (8th ed.). Englewood Cliffs, NJ: Prentice-Hall.

McIntyre, R., D. Smith, and C. Hassett. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69(1), 147-156.

Mitchell, S. K. (1979). Interrater agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Myers, D. H. and J. Heller. (1995). The dual role of AT&T's self-assessment process. Quality Progress, 28(1), 79 - 83.

National Institute of Standards and Technology. (1989). Malcolm Baldrige National Quality Award: 1990 application guidelines. Gaithersburg, MD.

National Institute of Standards and Technology. (1990). Malcolm Baldrige National Quality Award 1990 examiner preparation course materials. Gaithersburg, MD.

National Institute of Standards and Technology. (1994a). Malcolm Baldrige National Quality Award: 1995 award criteria. Gaithersburg, MD.

National Institute of Standards and Technology. (1994b). Malcolm Baldrige National Quality Award: 1995 examiner application. Gaithersburg, MD.

National Institute of Standards and Technology. (1994c). Malcolm Baldrige National Quality Award: 1994 Great Northern case study. Gaithersburg, MD.

National Institute of Standards and Technology. (1995a). Malcolm Baldrige National Quality Award: 1995 handbook for the board of examiners. Gaithersburg, MD.

National Institute of Standards and Technology. (1995b). Malcolm Baldrige National Quality Award: Case study packet, executive summary. Gaithersburg, MD.

National Institute of Standards and Technology. (1995c). Malcolm Baldrige National Quality Award: 1995 Colony Fasteners case study. Gaithersburg, MD.

National Institute of Standards and Technology. (1995d). Malcolm Baldrige National Quality Award: 1995 application scorebook. Gaithersburg, MD.

Newman, J. and G. Edgar (1990, December). Managing self assessment in the regulatory environment. Paper presented at the Executive Conference on Self Assessment for Self Assurance, American Nuclear Society, San Diego, CA.

Ott, L. (1984). An introduction to statistical methods and data analysis, Second Edition. Boston: Duxbury Press.

Pape, E. (1995). Personal correspondence with Garry Coleman, August 21.

Parrish, E. A. (1994). Report of the ABET criteria workshop, May 21-22, 1994. Unpublished manuscript.

Pearson, E. S. and Hartley, H. O. (Eds.). (1970). Biometrika tables for statisticians, Volume I (3rd ed. reprinted with additions). Cambridge: University Press.

Petersen, D. G. (1979). Accrediting Standards and Guidelines. Washington: Council on Postsecondary Accreditation.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decision Processes, 38, 76-91.

Pyzdek, T. (1995). "The Malcolm Baldrige National (Quality?) Award." Quality Progress, 28(5), 6.

Reimann, C. W. (1992). The Malcolm Baldrige National Quality Award. Hearing before the Subcommittee on Technology and Competitiveness of the Committee on Science, Space, and Technology, U. S. House of Representatives, February 5, (No. 97). Washington, DC: U. S. Government Printing Office, pp. 14-29.

Ritter, D. (1993). A tool for improvement using the Baldrige criteria. National Productivity Review, 12(2), 167-182.

Saal, Downey, & Lahey. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Sanford, R. L. (1993). Baxter Healthcare uses its own quality award to help achieve excellence. National Productivity Review, 12(1), pp. 37-43.

Schulman, R. (1994). Statistics for Social Science Research II, course note package, Blacksburg, Virginia: Virginia Tech.

Schwab, D. P., Heneman, H. G., III, and DeCotiis, T. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Selden, W. K. and H. V. Porter. (1977). Accreditation: Its purposes and uses. (An occasional paper). Washington: The Council on Postsecondary Accreditation.

Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Review, 11 (January), 22-40.

Smith, M. E. (1994). The NYNEX quality assessment plan. The use of quality award criteria and models for self-assessment purposes, Proceedings of the First European Forum on Quality Self-Assessment (pp. 181-198). Torino, Italy: European Organization for Quality.

Stamoulis, D. T. and N. Hauenstein. (1993). Rater training and rater accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78 (December), 994-1003.

Sulsky, L. M. and W. K. Balzer. (1988). Meaning and measurement of performance rating accuracy: some methodological and theoretical concerns. Journal of Applied Psychology, 73(3), 497-506.

Talley, D. J. (1989). The United States national contractor accreditation system. IEEE conference proceedings. (pp. 1624 - 1632).

Wachniak, R. (1990). An insider's view of the Malcolm Baldrige National Quality Award, presentation to the National Productivity Network, Chicago, IL, May 1-2.

Webb, N. M., G. L. Rowley, and R. J. Shavelson. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, pp. 81-90.

Young, K. E. (1983a). Accreditation: Complex evaluative tool. In K. E. Young, C. M. Chambers, H. R. Kells, and Associates. Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality. (pp. 17-35). San Francisco: Jossey-Bass.

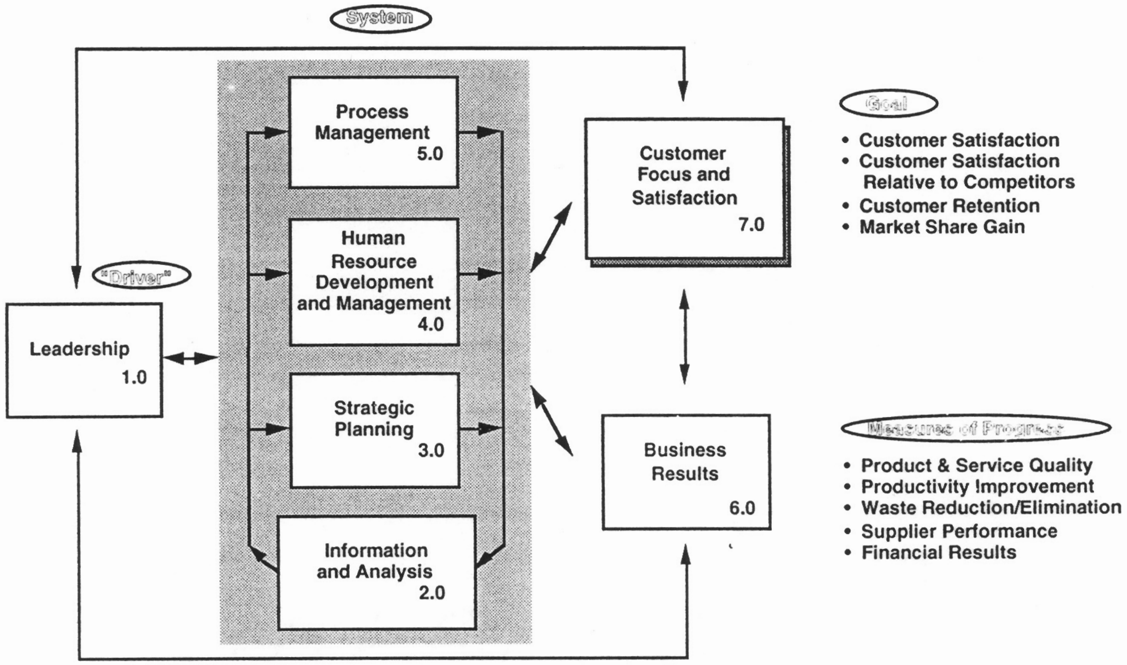
Young, K. E. (1983b). The changing scope of accreditation. In K. E. Young, C. M. Chambers, H. R. Kells, and Associates. Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality. (pp. 1-16). San Francisco: Jossey-Bass.

Young, K. E. (1983c). The future of accreditation. In K. E. Young, C. M. Chambers, H. R. Kells, and Associates. _Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality_. (pp. 379-406). San Francisco: Jossey-Bass.

Young, K. E., C. M. Chambers, H. R. Kells, and Associates. (1983). _Understanding accreditation: Contemporary perspectives on issues and practices in evaluating educational quality_. San Francisco: Jossey-Bass.

Appendix A. Baldrige Award Criteria Framework (NIST, 1994a)

BALDRIGE AWARD CRITERIA FRAMEWORK
Dynamic Relationships


System
"Driver"
Leadership 1.0
Process Management 5.0
Human Resource Development and Management 4.0
Strategic Planning 3.0
Information and Analysis 2.0
Customer Focus and Satisfaction 7.0
Business Results 6.0
Goal
• Customer Satisfaction
• Customer Satisfaction Relative to Competitors
• Customer Retention
• Market Share Gain
Measures of Progress
• Product & Service Quality
• Productivity !mprovement
• Waste Reduction/Elimination
• Supplier Performance
• Financial Results

Appendix B. 1995 Award Examination Criteria - Item Listing (NIST, 1994a)

1995 Examination Categories/Items		Point Values
1.0 Leadership		90
1.1 Senior Executive Leadership	45	
1.2 Leadership System and Organization	25	
1.3 Public Responsibility and Corporate Citizenship	20	
2.0 Information and Analysis		75
2.1 Management of Information and Data	20	
2.2 Competitive Comparisons and Benchmarking	15	
2.3 Analysis and Use of Company-Level Data	40	
3.0 Strategic Planning		55
3.1 Strategy Development	35	
3.2 Strategy Deployment	20	
4.0 Human Resource Development and Management		140
4.1 Human Resource Planning and Evaluation	20	
4.2 High Performance Work Systems	45	
4.3 Employee Education, Training, and Development	50	
4.4 Employee Well-Being and Satisfaction	25	
5.0 Process Management		140
5.1 Design and Introduction of Products and Services	40	
5.2 Process Management: Product and Service Production and Delivery	40	
5.3 Process Management: Support Services	30	
5.4 Management of Supplier Performance	30	
6.0 Business Results		250
6.1 Product and Service Quality Results	75	
6.2 Company Operational and Financial Results	130	
6.3 Supplier Performance Results	45	
7.0 Customer Focus and Satisfaction		250
7.1 Customer and Market Knowledge	30	
7.2 Customer Relationship Management	30	
7.3 Customer Satisfaction Determination	30	
7.4 Customer Satisfaction Results	100	
7.5 Customer Satisfaction Comparison	60	
TOTAL POINTS		1000

Appendix C. The Evaluation Process in a Nutshell (NIST, 1995a)

5.0 THE EVALUATION PROCESS

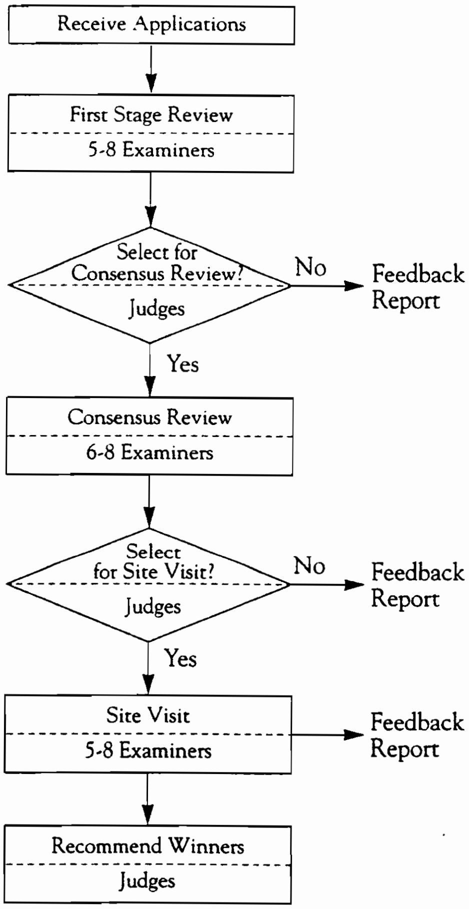
THE EVALUATION PROCESS IN A NUTSHELL

Each written application for the Malcolm Baldrige National Quality Award is evaluated by members of the Board of Examiners. High-scoring applicants are selected for site visits. Award recipients are chosen from among the applicants site-visited. All applicants receive a written feedback summary of Strengths and Areas for Improvement in their quality management.

KEY PROCESS STEPS

The following diagram illustrates the steps in the four-stage review:

Evaluation Process



Appendix D. Scoring Guidelines (NIST, 1994a)

SCORING GUIDELINES

SCORE	APPROACH/DEPLOYMENT
0%	• no systematic approach evident; anecdotal information
10% to 30%	• beginning of a systematic approach to the primary purposes of the Item • early stages of a transition from reacting to problems to a general improvement orientation • major gaps exist in deployment that would inhibit progress in achieving the primary purposes of the Item
40% to 60%	• a sound, systematic approach, responsive to the primary purposes of the Item • a fact-based improvement process in place in key areas; more emphasis is placed on improvement than on reaction to problems • no major gaps in deployment, though some areas or work units may be in very early stages of deployment
70% to 90%	• a sound, systematic approach, responsive to the overall purposes of the Item • a fact-based improvement process is a key management tool; clear evidence of refinement and improved integration as a result of improvement cycles and analysis • approach is well-deployed, with no major gaps; deployment may vary in some areas or work units
100%	• a sound, systematic approach, fully responsive to the overall purposes of the item • a very strong, fact-based improvement process is a key management tool; strong refinement and integration - backed by excellent analysis. • approach is fully deployed without any significant weaknesses or gaps in any areas or work units

SCORING GUIDELINES CONTINUED

SCORE	RESULTS
0%	• no results or poor results in areas reported
10% to 30%	• early stages of developing trends; some improvements *and/or* early good performance levels in a few areas • results not reported for many to most areas of importance to the applicant's key business requirements.
40% to 60%	• improvement trends *and/or* good performance levels reported for many to most areas of importance to the applicant's key business requirements • no pattern of adverse trends *and/or* poor performance levels in areas of importance to the applicant's key business requirements • some trends and/or current performance levels - evaluated against relevant comparisons *and/or* benchmarks - show areas of strength *and/or* good to very good relative performance levels
70% to 90%	• current performance is good to excellent in most areas of importance to the applicant's key business requirements • most improvement trends *and/or* performance levels are sustained • many to most trends *and/or* current performance levels - evaluated against relevant comparisons *and/or* benchmarks - show areas of leadership and very good relative performance levels
100%	• current performance is excellent in most areas of importance to the applicant's key business requirements • excellent improvement trends *and/or* sustained excellent performance levels in most areas • strong evidence of industry and benchmark leadership demonstrated in many areas

Appendix E. Training Plan (MBNQA Lecture(s) for Courses #4 and #5)

Objectives: The student will be able to:

- Describe the purposes of the MBNQA;
- Describe each of the seven categories of the MBNQA in a way that reflects the core values and concepts of the Award;
- Explain the relationships between the criteria in terms of the Award Criteria Framework (model);
- Describe how the Award Criteria may be used as part of an organization's improvement effort;
- Describe the three evaluation dimensions of approach, deployment, and results;
- Explain how the scoring guidelines for approach/deployment and results are applied to examination items.
- Use the criteria and scoring guidelines to assign a score to a given application (or category within), and identify strengths and areas for improvement that support the score, and identify site visit issues;
- Describe accuracy and consistency as they apply to scoring applications and how having a common frame of reference can improve accuracy and consistency.

Announcements:

- Collect second evaluation scorebooks from control group.

Reading Assignments:

- Award Criteria booklet (prior to initial evaluations);

Handouts/Problem Assignment:

- Selected vignettes from the Great Northern case study (in-class examples)
- Scoring worksheets for Great Northern vignettes
- Second evaluation scorebooks for treatment group

Topical Outline

- History and background of the MBNQA
- Overview of the Award process
- Award Criteria Framework and Characteristics
- Content of the categories
- Uses of the Award Criteria (ASQC survey)
- Self-assessment and self-regulation
- Accuracy and consistency of scores
- Evaluating an organization using the Award Criteria
  - + Approach, Deployment, and Results
  - + Relevance and importance to the business
  - + Scoring an item (discussion of examples)
  - + Strengths and areas for improvement
  - + Site visit issues
- Great Northern Case Study
  - + Business Overview
  - + Evaluation of selected items (sub-categories)
  - + Review of expert scoring
  - + Discussion of justification for expert scores
- Summary of MBNQA and this research project

Appendix F. Grading Procedures for the MBNQA Case Study

Each student evaluated four of seven categories from the Colony Fasteners Case Study. The case study represented an application for the MBNQA. Students were randomly assigned to categories. Each student was given a complete copy of the case study application and a scorebook for their assigned categories. Each student's completed evaluation (scorebook) was expected to contain at least the following parts:

- Cover page (name, # of hours worked, date completed)
- Key business factors worksheet
- Comment and scoring worksheets for each item (for two categories)
- Demographic and attitudinal questionnaire (only for second evaluation)
- A score summary worksheet

- A summary comment worksheet (the student was to keep these for preparing a final summary comment memo)

Following the training, each student submitted a final summary comment memo (approximately two pages). This memo addressed: overall evaluation of Colony Fasteners, lessons learned from this case study evaluation, how this case study evaluation fit into the overall course, and recommendations for improving the MBNQA training and case study evaluation.

Each student was graded on the following dimensions. Raw scores were calculated based on the weightings listed after each dimension.

- Thoroughness of identifying key business factors (compared to experts) - 5%
- Accuracy of scores (grade ranges established on boxplot of scores, based on expert scores) - 30%
- Thoroughness of identifying strengths and areas for improvement (compared to experts) - 35%
- Thoroughness of identifying site visit issues (compared to experts) - 10%
- Overall quality of final summary comment memo. - 20%

Each class was graded independently. Normative grading scales were developed for each group (treatment and control) to compensate for the benefit of the training (treatment group). Independent norms were established for the first and second evaluations (i.e., pre- and post-training). For the control group's second evaluation, each student's raw score was converted to a t-value based on the mean and standard deviation of their group for the second evaluation. The t-values were then converted to a percentage score based on the treatment group's mean and standard deviation for the second evaluation. A total average score was calculated for each student on a 0 to 100% scale by averaging their (adjusted) percentage scores on the first and second evaluations. Descriptive statistics (mean, median, standard deviation) were calculated for each class and each group. Along with the return of their graded scorebooks, students were given a memo that described how grades were determined and summarized the descriptive statistics.

Appendix G. Research Project - Work Breakdown Structure

I. Preparation
  A. Obtain materials
    1. MBNQA 1995 Award Criteria booklets
      a. Obtain mailing addresses for students in course #4 and course #5
      b. Submit consolidated request to NIST
      c. Students receive individual copies from NIST
        1. Confirm receipt via class announcement
        2. Provide copies to those who did not receive a copy from NIST
      d. Distribute copies from bulk shipment as needed
    2. Great Northern Case Study and Colony Fasteners Case Study
      a. Request from ASQC (Colony Fasteners not available until mid-Sept.)
      b. Use Great Northern to prepare work plan and research proposal
      c. Obtain Colony Fasteners in time to use in place of Great Northern
      d. Duplicate Colony Fasteners for distribution
  B. Develop sampling plan
    1. Randomly assign subjects to treatment or control group
    2. Determine how many and which categories to assign to subjects for evaluation
      a. For the first (pre-training) evaluation
      b. For the second (post-training) evaluation
    3. Develop and implement procedure for randomly assigning subjects to categories
    4. Identify evaluator characteristics to be measured
  C. Develop data collection forms and instructions
    1. Develop structure for data collection package (scorebooks)
      a. Develop cover sheet and determine contents for scorebooks
      b. Modify the worksheets from MBNQA Application Scorebook
      c. Incorporate questionnaire from task I.C.2. (second evaluation only)
      d. Write instructions for completing worksheets and questionnaire
      e. Develop coding scheme to track responses
    2. Develop questionnaire for measuring evaluator characteristics
    3. Pilot test questionnaire and scorebook instructions with previous students of course #5
    4. Revise materials per pilot test recommendations
  D. Obtain approval for use of human subjects
    1. Contact department representative to the Internal Review Board for use of human subjects
      a. Obtain forms to request approval for use of human subjects
      b. Solicit advice on preparation of request
    2. Prepare request for use of human subjects (including informed consent statement)
    3. Submit request and obtain approval
    4. Distribute informed consent statement to students
      a. Attend class to announce study purpose or prepare statement for instructor
      b. Ask students to read and sign statement
    5. Collect informed consent statements
  E. Write research proposal
    1. Meet with each committee member
      a. Share two page research alternative
      b. Describe study and solicit feedback
      c. Consult with committee members as needed
      d. Schedule proposal meeting
    2. Draft each section
      a. Introduction, context, and research problem
      b. Literature review
      c. Research methodology
        1. Research questions and hypotheses

2. Data to be collected and analyzed

3. Description of work plan

d. Bibliography

e. Appendices: example tools

3. Review each section with committee chair

4. Compile and reproduce complete document

a. Prepare cover page, TOC, formatting, etc.

b. Duplicate and bind

5. Send complete copy to each committee member (cc: instructors)

6. Continue literature review while committee members read

7. Hold proposal meeting with committee

a. Collect feedback and suggestions

b. Obtain committee approval and signatures

8. Finalize work plan with instructors

II. Pre-Training Data Collection

A. Announce assignment in each class

1. Confirm each subject has the MBNQA criteria booklet, assign reading

2. Announce assignment one week before distributing packets

B. Distribute case studies and scorebooks (treatment and control groups)

1. Follow results of randomization procedure when distributing scorebooks

2. Prepare instructions for off-campus site coordinators

3. Prepare masters and copies for mailing and distribution

4. Mail copies to off-campus sites with regular class materials

5. Distribute copies in-class to on-campus subjects

C. Collect completed scorebooks one week after distributing (treatment and control groups)

1. Have off-campus subjects submit scorebooks as they do regular assignments

2. Copy comment and scoring worksheets for later analysis

a. Provide instructors with expert evaluations of the case study

b. Grade scorebooks

c. Delay return of graded materials until after training and second evaluations

3. Compile worksheets by categories to prepare for data entry

D. Track responses received

1. Maintain a log of scorebooks, questionnaires, and memos received

2. Call, e-mail, or write to solicit additional responses if necessary

E. Distribute scorebooks and questionnaires for second evaluation (control group only)

1. Follow results of randomization procedure when distributing scorebooks and questionnaires

2. Prepare instructions for off-campus site coordinators

3. Prepare masters and copies for mailing and distribution

4. Mail copies to off-campus sites with regular class materials

5. Distribute copies in-class to on-campus subjects

F. Collect completed scorebooks and questionnaires one week after distributing (control group only)

1. Have off-campus subjects submit scorebooks and questionnaires as they do regular assignments

2. Copy comment and scoring worksheets for later analysis

a. Grade scorebooks

b. Delay return of graded materials until after training and second evaluations

3. Compile worksheets by categories to prepare for data entry

G. Confirm responses received from all consenting subjects (control group)

1. Call, e-mail, or write to solicit additional responses if necessary

2. Discard copied worksheets from non-consenting subjects

III. Training
   A. Develop student objectives for a 150 minute training module
      1. Review objectives with instructors
      2. Modify objectives as required
      3. Develop topical outline
   B. Develop materials for training module
      1. Announce the MBNQA criteria booklets will be used as text during training
      2. Prepare handouts to be distributed in advance
         a. Application vignettes
         b. Scoring worksheets
      3. Prepare visuals (overheads) for televised presentation
      4. Practice training module to assess and adjust length
   C. Distribute training materials
      1. Prepare instructions for off-campus site coordinators
      2. Prepare masters and copies for mailing and distribution
      3. Mail copies to off-campus sites with regular class materials
      4. Distribute copies in-class to on-campus subjects
   D. Deliver training
      1. Course #4 on a Thursday (meets once per week for 150 minutes)
      2. Course #5 on Monday and Wednesday (meets twice per week for 75 minutes each)
   E. Announce and distribute post-training assignments at end of training

IV. Post-Training Data Collection
   A. Distribute scorebooks and questionnaires (treatment group only)
      1. Follow results of randomization procedure when distributing scorebooks and questionnaires
      2. Prepare instructions for off-campus site coordinators
      3. Prepare masters and copies for mailing and distribution
      4. Mail copies to off-campus sites with regular class materials
      5. Distribute copies in-class to on-campus subjects
   B. Assign final summary comment memos (due in one week)
   C. Collect completed scorebooks and questionnaires (treatment group only) and final summary comment memos (all subjects) one week after distributing/assigning
      1. Have off-campus subjects submit packages as they do regular assignments
      2. Copy comment and scoring worksheets for later analysis
         a. Grade scorebooks
         b. Return graded materials from pre-training data collection (after task IV.C.)
         c. Return graded post-training materials
      3. Compile worksheets by categories to prepare for data entry
   D. Track responses received
      1. Maintain a log of scorebooks, questionnaires, and memos received
      2. Call, e-mail, or write to solicit additional responses if necessary

V. Data analysis
   A. Develop data collection templates
      1. Excel and Minitab spreadsheets for quantitative data
      2. Microsoft Word tables for qualitative data
   B. Enter data
      1. Enter scores for each item
      2. Enter evaluator characteristics data
      3. Enter strengths, areas for improvement, and site visit issues
      4. Proof and edit data entry
   C. Perform data analyses (see structure of the research problem)
      1. Perform data analysis procedures
      2. Evaluate and interpret results
      3. Write interpretation and portray results

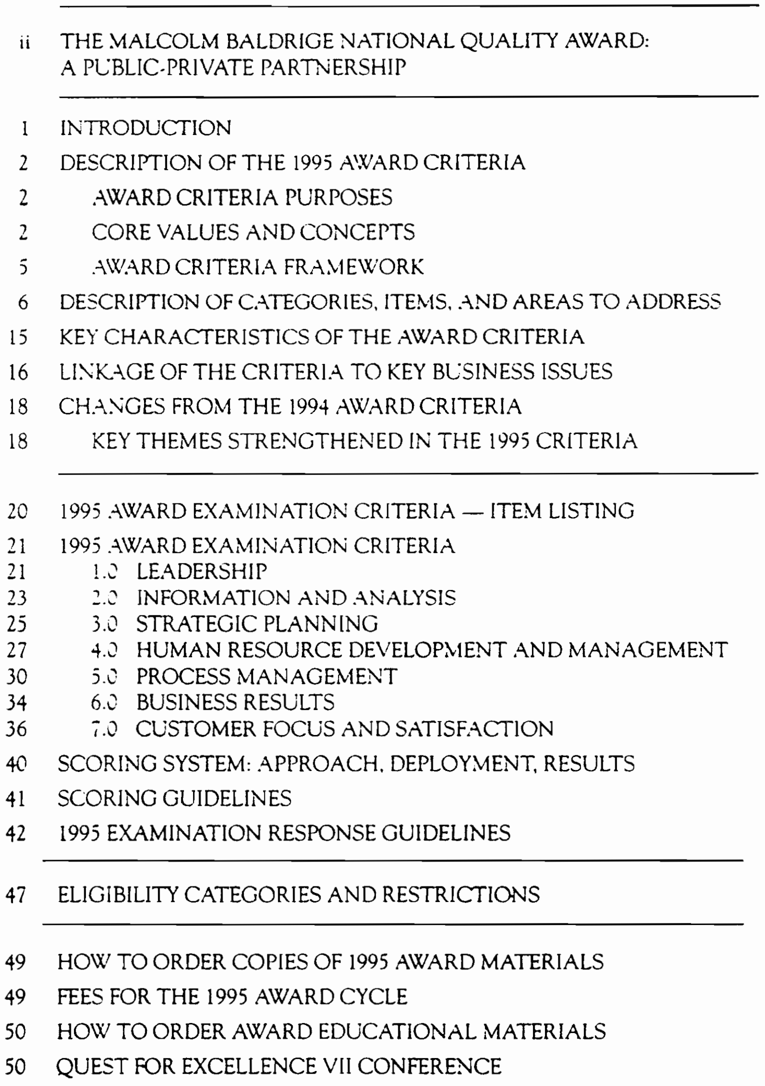
VI. Communication of results

A. Write research report (dissertation)

1. Develop table of contents
2. Draft each section
3. Review each section with committee chair
   a. Revise per feedback
   b. Meet with individual committee members as needed
3. Prepare complete document
   a. Integrate across sections
   b. Prepare figures, tables, appendices
   c. Ensure proper formatting and page numbering
4. Compile and reproduce complete document
   a. Print master copy
   b. Duplicate and bind
5. Send complete copy to each committee member (cc: instructors)
6. Hold defense meeting with committee
   a. Schedule defense meeting
   b. Obtain forms from graduate school
   c. Collect feedback and suggestions
   d. Obtain committee approval and signatures
7. Make final revisions and submit to graduate school

B. Publish paper(s) summarizing one or more aspects of the project

Appendix H. Contents - MBNQA 1995 Award Criteria (NIST, 1994a)



ii THE MALCOLM BALDRIGE NATIONAL QUALITY AWARD:
A PUBLIC-PRIVATE PARTNERSHIP

1 INTRODUCTION
2 DESCRIPTION OF THE 1995 AWARD CRITERIA
2 AWARD CRITERIA PURPOSES
2 CORE VALUES AND CONCEPTS
5 AWARD CRITERIA FRAMEWORK
6 DESCRIPTION OF CATEGORIES, ITEMS, AND AREAS TO ADDRESS
15 KEY CHARACTERISTICS OF THE AWARD CRITERIA
16 LINKAGE OF THE CRITERIA TO KEY BUSINESS ISSUES
18 CHANGES FROM THE 1994 AWARD CRITERIA
18 KEY THEMES STRENGTHENED IN THE 1995 CRITERIA

20 1995 AWARD EXAMINATION CRITERIA — ITEM LISTING
21 1995 AWARD EXAMINATION CRITERIA
21 1.0 LEADERSHIP
23 2.0 INFORMATION AND ANALYSIS
25 3.0 STRATEGIC PLANNING
27 4.0 HUMAN RESOURCE DEVELOPMENT AND MANAGEMENT
30 5.0 PROCESS MANAGEMENT
34 6.0 BUSINESS RESULTS
36 7.0 CUSTOMER FOCUS AND SATISFACTION
40 SCORING SYSTEM: APPROACH, DEPLOYMENT, RESULTS
41 SCORING GUIDELINES
42 1995 EXAMINATION RESPONSE GUIDELINES

47 ELIGIBILITY CATEGORIES AND RESTRICTIONS

49 HOW TO ORDER COPIES OF 1995 AWARD MATERIALS
49 FEES FOR THE 1995 AWARD CYCLE
50 HOW TO ORDER AWARD EDUCATIONAL MATERIALS
50 QUEST FOR EXCELLENCE VII CONFERENCE

Appendix 1. Contents - Colony Fasteners Case Study (NIST, 1995c)

CASE STUDY

TABLE OF CONTENTS

Appendix J. Key Business Factors Worksheet (NIST, 1995d)

Key Business Factors Worksheet

To begin the scoring process, review the Application Overview and list the Key Business Factors for this applicant. For a discussion of the Guidelines for Preparing the Business Overview, see page 42 of the *1995 Award Criteria* book.

Appendix K. Example Comment and Scoring Worksheet (NIST, 1995d)

Comment and Scoring Worksheet

1.0 Leadership (90 pts.)

The *Leadership* Category examines senior executives' personal leadership and involvement in creating and sustaining a customer focus, clear values and expectations, and a leadership system that promotes performance excellence. Also examined is how the values and expectations are integrated into the company's management system, including how the company addresses its public responsibilities and corporate citizenship.

1.1 Senior Executive Leadership *(45 pts.)*

Describe senior executives' leadership and personal involvement in setting directions and in developing and maintaining a leadership system for performance excellence.

A D R

☑—☑ □

AREAS TO ADDRESS

a. how senior executives provide effective leadership and direction in building and improving company competitiveness, performance, and capabilities. Describe executives' roles in: (1) creating and reinforcing values and expectations throughout the company's leadership system; (2) setting directions and performance excellence goals through strategic and business planning; and (3) reviewing overall company performance, including customer-related and operational performance.

b. how senior executives evaluate and improve the effectiveness of the company's leadership system and organization to pursue performance excellence goals.

(See page 40 of the *1995 Award Criteria* for a description of these symbols.)

Notes:

(1) "Senior executives" means the applicant's highest-ranking official and those reporting directly to that official.

(2) Values and expectations [1.1a(1)] should take into account all stakeholders – customers, employees, stockholders, suppliers and partners, the community, and the public.

(3) Activities of senior executives appropriate for inclusion in 1.1a might also include customer, employee, and supplier interactions, mentoring other executives, benchmarking, and employee recognition.

(4) Review of company performance is addressed in 1.2c. Responses to 1.1a(3) should reflect senior executives' personal leadership of and involvement in such reviews, and their use of the reviews to focus on key business objectives.

(5) Evaluation of the company's leadership system might include assessment of executives by peers, direct reports, or a board of directors. It might also include results of surveys of company employees.

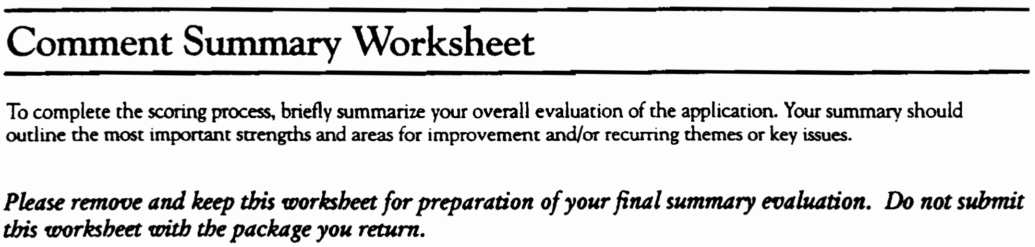
Appendix K. Example Comment and Scoring Worksheet Continued

1.1 Senior Executive Leadership (45 points)	EXAMINER INITIALS □	PERCENT SCORE □

+/++	Area to Address	(+) STRENGTHS
−/− −	Area to Address	(−) AREAS FOR IMPROVEMENT

SITE VISIT ISSUES:

Appendix L. Comment Summary Worksheet (adapted from NIST, 1995d)



Comment Summary Worksheet

To complete the scoring process, briefly summarize your overall evaluation of the application. Your summary should outline the most important strengths and areas for improvement and/or recurring themes or key issues.

Please remove and keep this worksheet for preparation of your final summary evaluation. Do not submit this worksheet with the package you return.

Appendix M. Score Summary Worksheet (NIST, 1995d)

Score Summary Worksheet

Examiner Name ____________________

Applicant Name COLONY FASTENERS ____________ Applicant Number ____________

SUMMARY OF EXAMINATION ITEMS	Total Points Possible A	Percent Score 0-100% (10% units) B	Score (A x B) C
1.0 LEADERSHIP 90 POSSIBLE POINTS			
1.1 Senior Executive Leadership	45	______ %	______
1.2 Leadership System and Organization	25	______ %	______
1.3 Public Responsibility and Corporate Citizenship	20	______ %	______
Category Total	90 SUM A		______ SUM C
2.0 INFORMATION AND ANALYSIS 75 POSSIBLE POINTS			
2.1 Management of Information and Data	20	______ %	______
2.2 Competitive Comparisons and Benchmarking	15	______ %	______
2.3 Analysis and Use of Company-Level Data	40	______ %	______
Category Total	75 SUM A		______ SUM C
3.0 STRATEGIC PLANNING 55 POSSIBLE POINTS			
3.1 Strategy Development	35	______ %	______
3.2 Strategy Deplovment	20	______ %	______
Category Total	55 SUM A		______ SUM C
4.0 HUMAN RESOURCE DEVELOPMENT AND MANAGEMENT 140 POSSIBLE POINTS			
4.1 Human Resource Planning and Evaluation	20	______ %	______
4.2 High Performance Work Systems	45	______ %	______
4.3 Employee Education, Training, and Development	50	______ %	______
4.4 Employee Well-Being and Satisfaction	25	______ %	______
Category Total	140 SUM A		______ SUM C
5.0 PROCESS MANAGEMENT 140 POSSIBLE POINTS			
5.1 Design and Introduction of Products and Services	40	______ %	______
5.2 Process Management: Product and Service Production and Delivery	40	______ %	______
5.3 Process Management: Support Services	30	______ %	______
5.4 Management of Supplier Performance	30	______ %	______
Category Total	140 SUM A		______ SUM C
6.0 BUSINESS RESULTS 250 POSSIBLE POINTS			
6.1 Product and Service Quality Results	75	______ %	______
6.2 Company Operational and Financial Results	130	______ %	______
6.3 Supplier Performance Results	45	______ %	______
Category Total	250 SUM A		______ SUM C
7.0 CUSTOMER FOCUS AND SATISFACTION 250 POSSIBLE POINTS			
7.1 Customer and Market Knowledge	30	______ %	______
7.2 Customer Relationship Management	30	______ %	______
7.3 Customer Satisfaction Determination	30	______ %	______
7.4 Customer Satisfaction Results	100	______ %	______
7.5 Customer Satisfaction Comparison	60	______ %	______
Category Total	250 SUM A		______ SUM C
GRAND TOTAL (D)	1000		______ D

Appendix N. Evaluator Questionnaire

This questionnaire will provide information about evaluator characteristics and opinions to be compared to the evaluators' scores of the Colony Fasteners Case Study. Please complete each question to the best of your knowledge. Please return this questionnaire along with your scorebook.

Part I. This section asks for information regarding your evaluation of the Colony Fasteners Case Study using the Malcolm Baldrige National Quality Award criteria.

1. Which four categories did you use to evaluate the case study?

_______ 1.0 Leadership
_______ 2.0 Information and Analysis
_______ 3.0 Strategic Planning
_______ 4.0 Human Resource Development and Management
_______ 5.0 Process Management
_______ 6.0 Business Results
_______ 7.0 Customer Focus and Satisfaction

2. For the categories you evaluated, please rate how difficult they were to evaluate using the following scale: easy = 1; somewhat easy = 2; somewhat difficult = 3; difficult = 4.

_______ 1.0 Leadership
_______ 2.0 Information and Analysis
_______ 3.0 Strategic Planning
_______ 4.0 Human Resource Development and Management
_______ 5.0 Process Management
_______ 6.0 Business Results
_______ 7.0 Customer Focus and Satisfaction

3. For the categories you evaluated, please rate how close to the experts' scores you feel your scores were using the following scale: remote = 1; somewhat remote = 2; somewhat close = 3; close = 4.

_______ 1.0 Leadership
_______ 2.0 Information and Analysis
_______ 3.0 Strategic Planning
_______ 4.0 Human Resource Development and Management
_______ 5.0 Process Management
_______ 6.0 Business Results
_______ 7.0 Customer Focus and Satisfaction

Part II. This section asks for information regarding your experience with the Malcolm Baldrige National Quality Award criteria and other forms of crganizational assessments.

4. Had you personally used the Malcolm Baldrige National Quality Award criteria for any purpose prior to receiving a copy of the criteria in this course?

_______ (1) Yes

_______ (2) No If you answered No to question 4, please skip to question 6.

5. Referring to your answer to question 4, how did you use the Malcolm Baldrige National Quality Award criteria? Please check all that apply.

_______ (1) To apply for an award (company, state, or national).

_______ (2) As a written self-assessment tool for an entire company.

_______ (3) As a written self-assessment tool for a department or division.

_______ (4) As an informal self-assessment tool for an entire company.

_______ (5) As an informal self-assessment tool for a department or division.

_______ (6) As a tool to set up quality processes in an entire company.

_______ (7) As a tool to set up quality processes in a department or division.

_______ (8) As a tool to improve existing quality processes in an entire company.

_______ (9) As a tool to improve existing quality processes in a department or division.

_______ (10) As a common language to communicate with others inside a company (e.g., departments or divisions).

_______ (11) As a common language to communicate with outside business partners (e.g., suppliers).

_______ (12) As a common language to communicate with other companies.

_______ (13) As a common language to communicate with public-sector institutions (e.g., education or health care organizations).

_______ (14) As part of the curriculum for a course, seminar, or workshop.

_______ (15) As a source of information on how to achieve business excellence.

6. In what other forms of organizational assessments have you participated? Please list. Examples include: ISO 9000 certification, compliance audits, supplier certification (e.g, Ford's Q1), National Contractor Accreditation, and quality or productivity award application. If none, simply write "none."

Part III. This section asks for information regarding your background and current professional activity.

7. What is the highest degree you have obtained?

_______ (1) Have not obtained Bachelor

_______ (2) Bachelor

_______ (3) Master

_______ (4) Doctorate

_______ (5) Graduate other than Master or Doctorate, please identify: ____________________

8. What was the educational specialty of your highest degree obtained?

_______ (1) Business

_______ (2) Engineering

_______ (3) Education

_______ (4) Math or Statistics

_______ (5) Physical Sciences (e.g., biology, chemistry, geology)

_______ (6) Social Sciences (e.g., psychology, sociology)

_______ (7) Other (please identify) ______________________________

9. In addition to your formal education, how much quality and productivity improvement training and development have you received in the past ten years? For each category, please estimate the number of days you attended.

(A) _________ Days of classroom training (e.g., short courses, seminars)

(B) _________ Days of on-the-job training

(C) _________ Days of conference sessions attended

10. How many years of professional work experience do you have? Please pro-rate part-time employment to reflect the proportion of a full-time position.

_________ years

11. What is your current job function? Please list below. Examples include: Executive/Administrative; Production/Service Delivery; Quality Control/Quality Assurance; Maintenance; Human Resources; Finance/Accounting; Sales/Marketing; Research & Development; Teaching/Training; Engineering; Full time student; Unemployed.

Job function: ____

12. How many years of work experience do you have in a quality control, quality assurance, or quality improvement function? If none, simply write "0."

_________ years in QC, QA, or QI

13. Which of the following statements best describes the supervisory responsibility of your current job?

_______ (1) I have no consistent supervisory responsibility.
If you chose item (1), please skip to question 15.

_______ (2) I supervise other employees as a regular part of my job.

14. How many employees do you supervise or supervise through subordinate supervisors in your current job?

_________ Number of employees supervised or directed.

15. Which of the following categories best describes your current employer?

_______ (1) Manufacturing

_______ (2) Service

_______ (3) Federal Government

_______ (4) State or Local Government

_______ (5) Education

_______ (6) Health Care

15. What is the total number of employees in your *entire* company or employing organization? Federal and state government employees should write "Govt." rather than a number in the space provided.

_________ total number of employees

16. What is your current age? __________

17. What is your gender? ________(1) Female ________(2) Male

Thank you very much for your participation in this study. Please return your questionnaire with your scorebook.

Appendix O. Informed Consent Statement

Title of Project: Estimating the impact of third-party reviewer training and characteristics on the scoring of written organizational self-assessments.

Principal Investigator: Garry Coleman

I. Purpose of this Research

You are invited to participate in a study about the effects of evaluator training and characteristics on the scoring of organizational self-assessments based on the Malcolm Baldrige National Quality Award criteria.

II. Procedures

As a regular part of course #4 or course #5, you will be provided the Malcolm Baldrige National Quality Award criteria and an organizational case study to evaluate based on these criteria. This study will use the raw data from the evaluations of the case study and compare these to evaluator characteristics. Data needed specifically for this study include students' educational background, work experience, exposure to the Baldrige criteria, etc. These data will be collected via a brief questionnaire. The questionnaire takes approximately ten minutes to complete.

III. Benefits of this project

Your participation in this project will provide data that currently isn't available, even though use of the Baldrige and other Award criteria for self-assessments is widespread.

IV. Extent of Confidentiality

All responses will be kept strictly confidential. At no time will the researchers release the results of the study to anyone other than individuals working on the project without your written consent. Individual responses will have the names removed and be coded for data analyses and any subsequent publication.

V. Compensation

No course credit or compensation is provided for completion of the questionnaire. Evaluation of the case study is a regular graded assignment in course #4 and course #5; however, your decision to participate or not participate in this study will have **absolutely no impact on your grade in the course.**

VI. Freedom to Withdraw

You are free to withdraw from this study at anytime by notifying the principal investigator.

VII. Approval of Research

This research project has been approved, as required, by the Institutional Review Board for projects involving human subjects at Virginia Tech, and by the Department of Industrial and Systems Engineering.

VIII. Subject's Permission

I have read and understand the informed consent and conditions of this project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project. If I participate, I may withdraw from this study at any time by notifying the principal investigator. I agree to abide by the rules of this project.

____

Signature

Should you have any questions about this research or its conduct, you may contact Dr. Bob Beaton (231-5936) or Dr. Ernest Stout (231-9359).

Appendix P. Course #4 - Announcement, 28 September, 1995

Course #4 - Announcement
28 September, 1995

Later this semester we will be studying the Malcolm Baldrige National Quality Award and evaluating a case study based on this award.

We have requested copies of the 1995 Baldrige Award Criteria for everyone registered in this class. A copy will be mailed directly to you, at the mailing address provided us by the University Registrar's Office. You should be receiving your copy soon, if you haven't already.

If you do not receive a copy of the 1995 Baldrige Award Criteria by Thursday, October 5, please contact Garry Coleman at the ISE office in Blacksburg (540) 231-XXXX or e-mail: xxxxx@yyy. He will check your address and a second copy will be sent to you immediately. Thank you.

Appendix Q. Data Collection and Training Schedule

Date	course #4 (Thursdays'	course #5 (Mon., Wed.)
H 10/12	Deliver first evaluation materials (case study and scorebooks) to distributor Ship first evaluation materials to off-campus site coordinators	
M 10/16	Deliver control group's second evaluation materials (scorebooks & questionnaire) to distributor	
H 10/19	Distribute first evaluation materials to all subjects Ship control group's second evaluation materials to off-campus site coordinators	Deliver first evaluation materials (case study and scorebooks) to distributor
M 10/23		Ship first evaluation materials to off-campus site coordinators
H 10/26	All subjects submit completed first evaluation scorebooks to instructor or site coordinator Site coordinators forward scorebooks to researcher Distribute second evaluation materials to control group only Ship treatment group's second evaluation materials and lecture notes for all to off-campus site coordinators	
M 10/30		Distribute first evaluation materials to all subjects Ship control group's second evaluation materials to off-campus site coordinators
H 11/2	Control group submits completed second evaluation scorebooks & questionnaires to instructor or site coordinator Site coordinators forward scorebooks & questionnaires to researcher MBNQA Lecture - distribute lecture notes Distribute treatment group's second evaluation materials	

Date	course #4 (Thursdays)	course #5 (Mon., Wed.)
M 11/6		All subjects submit completed first evaluation scorebooks to instructor or site coordinator Site coordinators forward scorebooks to researcher Distribute second evaluation materials to control group only Ship treatment group's second evaluation materials and lecture notes for all to off-campus site coordinators
H 11/9	Treatment group submits completed second evaluation scorebooks & questionnaires to instructor or site coordinator All subjects submit final summary comment memos Site coordinators forward scorebooks, questionnaires, and final memos to researcher	
M 11/13 & W 11/15		Control group submits completed second evaluation scorebooks & questionnaires to instructor or site coordinator Site coordinators forward scorebooks & questionnaires to researcher MBNQA Lecture - distribute lecture notes Distribute treatment group's second evaluation materials (11/15)
11/27		Treatment group submits completed second evaluation scorebooks & questionnaires to instructor or site coordinator All subjects submit final summary comment memos Site coordinators forward scorebooks, questionnaires, and final memos to researcher OR Treatment group mails completed second evaluation scorebooks and questionnaires to researcher postmarked nlt 11/22 and final memos are collected on 11/27

Appendix R. Memorandum of Agreement

The purpose of this memorandum of agreement (MOA) is to describe how Garry Coleman, the researcher, will work with __________, instructor of course #5, and __________, instructor of course #4, to collect data for his research in the context of course #5 and course #4.

The researcher will serve as a guest lecturer for each course. He will lecture on the Malcolm Baldrige National Quality Award and its uses for organizational assessment and improvement. He will collaborate with the instructors and students to obtain a free copy of the Malcolm Baldrige National Quality Award criteria booklet for each student. He will provide a case study for the students to evaluate as a graded course assignment. He will provide a case study scorebook to structure the evaluations. He will provide each instructor with a key (expert evaluation) to the case study evaluation.

The researcher will grade the students' evaluations of the case study and provide a listing of the grades to the appropriate instructor. The graded case study scorebooks will be returned to the students. The students' grades on this assignment will not be used for research purposes. A student's decision to not participate in the research study will not affect their grade.

The researcher will use the students' evaluations of the case study as data for his research. He will also ask each student to complete a brief questionnaire describing the subject's demographic characteristics and his or her perceptions regarding the evaluation of the case study. The data from this questionnaire will be used for research purposes only and will not affect the student's grade in the course. The case study evaluations and questionnaire data will be kept strictly confidential.

Both course #5 and course #4 are televised courses. The researcher will prepare materials for distribution. The instructors will provide a roster of students and their class locations to facilitate the packaging of materials. The instructors will distribute and collect the lecture, case study evaluation assignment, informed consent statements, and questionnaire materials via their usual televised course procedures. The instructors will make announcements or provide the researcher the opportunity to make announcements as required for the administration of the case study evaluations. The researcher will provide written announcements for the instructors if required.

The guest lecture and case study evaluation assignment are designed to support the objectives of both courses. Only the demographic questionnaire of subjects is outside of the course objectives.

Advance planning and scheduling are necessary to facilitate an effective lecture, case study evaluation assignment, and questionnaire. The following documents are attached to clarify the details of the lecture, assignment, and research project.

- Data Collection and Training Schedule
- Training Plan (for guest lecture)
- Grading Plan for Case Study
- Work Breakdown Structure for the Research Project

This collaboration represents a win-win situation for the students, instructors, and researcher.

______________________________ ______________________________

Garry Coleman Instructor, Course #5

____

Instructor, Course #4

Appendix S. Assignment Procedures

Assignment of Subjects to Groups (10/4/95)

All subjects were listed in alphabetical order. A statistics text (Ott) was opened to a page in the middle of the book and the last two digits of the page number were used to enter a random number table (first digit = row #, second digit = column #). The first two digits in each block of five digits were read and used to locate the XYth student on the alphabetical listing. The first forty-nine original pairs of such digits were used to assign the students to the control group. The remaining students were assigned to the treatment group. Students taking both[1] course #4 and course #5 were assigned only once. This procedure resulted in 49 control and 49 treatment group subjects.

For grading purposes, the students taking both classes will be treated as part of the treatment group in course #5 due to learning from course #4 (where they will be treated as part of their assigned group).

A coding scheme was developed to identify subjects. A four digit code was assigned to each subject. The first digit represented group (1 = control; 2 = treatment). The second digit represented course (#4 or #5). The third and fourth digits were the sequential number assigned the subject from the list of subjects by class sorted alphabetically by site.

Assignment of Subjects to Categories for the First Evaluation

Subjects were assigned to categories by group for the first evaluation, starting with the control group. Subjects were numbered from 1 to 49 in the alphabetical listing of the group. Within each group, fourteen subjects were assigned to each category. Starting with category one, a random number table was entered as before. The last two digits in each block of digits were used to identify which of the group's subjects were to be assigned to that category. The digits were read until fourteen subjects had been assigned, then assignment to the next category began (sequence of categories: 1, 2, 4, 6, 7, 3, 5). Once a subject had been assigned to two categories, they were no longer eligible. When only two categories remained to be assigned, any subjects with no assigned categories were assigned to both of the remaining categories. Then the remaining assignments were made. This procedure was repeated for the treatment group; however, the sequence of categories was reordered using a random number table (sequence of categories: 6, 7, 3, 4, 1, 2, 5).

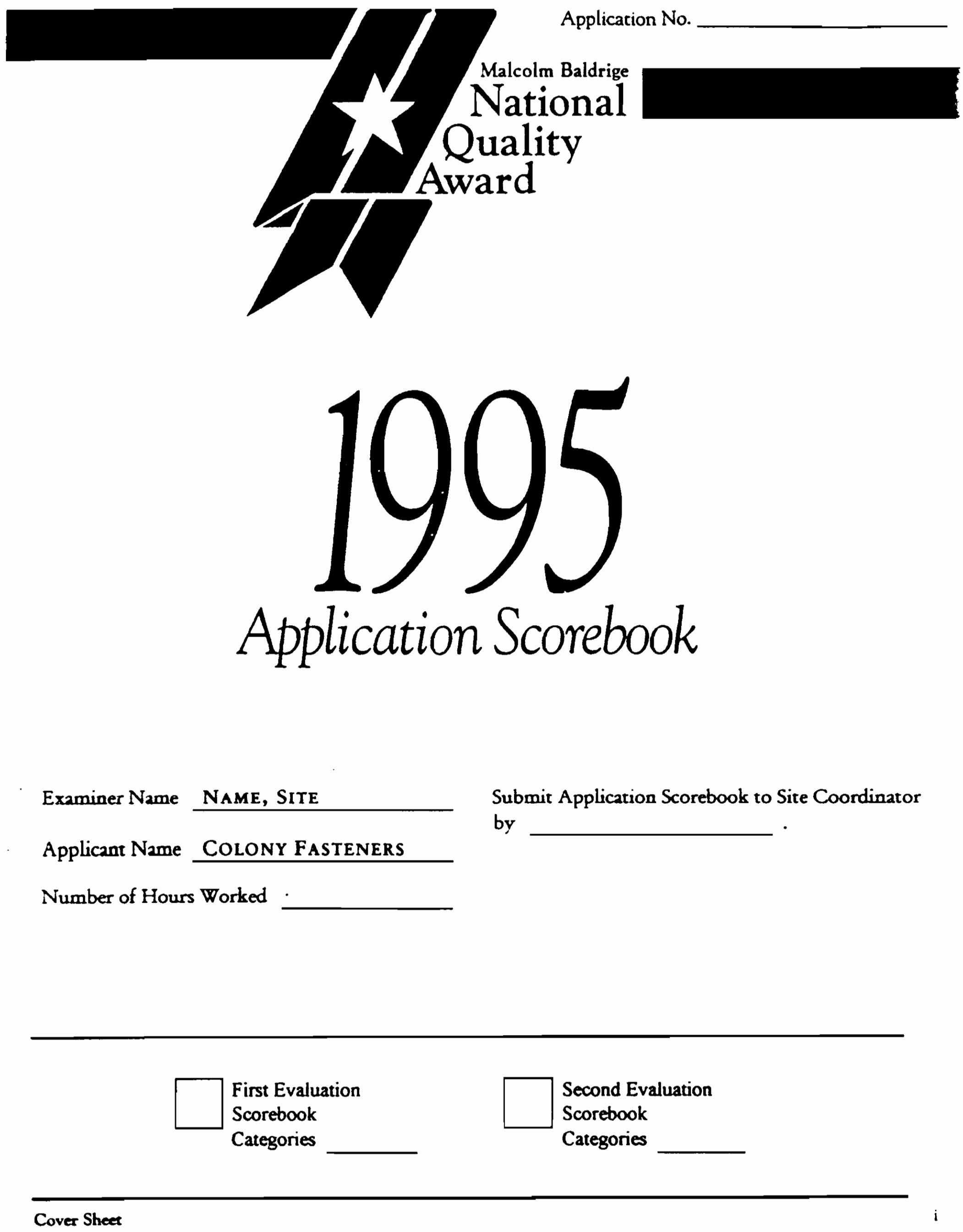
[1] (Revised 1/22/96) All of the results from these "double" students assigned to the control group in course #4 were used for experimental purposes The results from the double students assigned to the treatment group in course #4 were used only for the first evaluation. The treatment group's second evaluation in course #4 were not used. Those assigned to the treatment group had completed their first evaluation in course #5 before completing the second evaluation in course #4. The results of all double students' evaluations from course #5 were not used for experimental purposes, due to potential learning from their previous evaluation experience in course #4.

Assignment of Subjects to Categories for the First Evaluation (10/9/95)

The assignment of subjects to categories for the second evaluation was also done by group. First, the subjects were sorted by four digit code[2], thus separating the two groups. Then the subjects were numbered sequentially for each group. Assignment of categories followed a similar procedure as before, with the added constraint that a subject could not be assigned to a category they had been assigned to for the first evaluation. The sequence of categories for assignment of the control group were: 1, 6, 7, 5, 4, 3, 2. The sequence of categories for assignment of the treatment group were: 5, 2, 3, 1, 4, 7, 6. The duplicate subjects in course #4 were assigned to categories via a random number table (Excel generated from 1 to 7) to produce a balanced assignment of two subjects per category. In the event they were omitted from the final analysis, they would not impact balance. The same duplicate subjects in course #5 were each assigned to the three remaining categories (those they had not been assigned to in course #4) and to a randomly assigned fourth category (using the Excel generated random number table).

[2] The students taking both classes were separated from this sorting. Assignment for these students was done by a separate randomization procedure, to reduce the impact of their possible omission on the balance of subjects per category. An adjustment was made to their assignments for the first evaluation to achieve the same effect. Basically, one subject's group and assignment were switched with a subject not in both classes.

Appendix T. Example Scorebook Cover Page (adapted from NIST, 1995d)



Application No. ____________________

Malcolm Baldrige
National
Quality
Award

1995
Application Scorebook

Examiner Name NAME, SITE

Applicant Name COLONY FASTENERS

Number of Hours Worked ____________

Submit Application Scorebook to Site Coordinator by ____________________ .

□ First Evaluation Scorebook Categories ________

□ Second Evaluation Scorebook Categories ________

Cover Sheet i

Appendix U. Example Cover Memo for 1st Evaluation

To: ______________________________________(Subject's Name)
Course #4

From: Garry D. Coleman, Instructor
Guest Lecturer for Course #4

Re: Colony Fasteners Case Study

Date: 19 October 1995

I will be giving a guest lecture on the Malcolm Baldrige National Quality Award on November 2. In conjunction with this lecture, you will evaluate and score a case study based on the Baldrige criteria. You should already have received a personal copy of the Baldrige criteria booklet. If not, your site coordinator has extra copies. Please confirm that you have also received the following items along with this memo:

- The Colony Fasteners Case Study,
- A 1995 Application Scorebook, and
- An Informed Consent Statement (attached, for participation in a related research project).

At this time, you are asked to evaluate and score the case study against two of the seven Baldrige criteria categories. *The evaluation and scoring of your first two randomly assigned categories is due on October 26, 1995.* The two categories you have been assigned are: ___________. These same category numbers are also listed on the cover of your Scorebook.

Instructions:

1. Please read the attached informed consent statement. Please signify your willingness to participate in this research project by signing the form and returning it to your site coordinator.
2. If you have not already done so, please read the 1995 Baldrige Award Criteria booklet. You should become familiar with the descriptions of the criteria (pp. 2-17); the award examination criteria (pp. 21-39); and the scoring system (pp. 40-41).
3. Read the Colony Fasteners Case Study. Assume Colony Fasteners has asked you to perform an assessment based on the Baldrige criteria. The only information you have about Colony Fasteners is the written case study, which is provided in the form of a mock application for the Baldrige Award.
4. Evaluate the case study against the two categories you have been assigned using the 1995 Application Scorebook. Please complete the following per the instructions in the Scorebook:
   - the Key Business Factors Worksheet,
   - the applicable Comment and Scoring Worksheets,
   - the Comment Summary Worksheet,
   - the Score Summary Worksheet, and
   - the Conflict of Interest Statement.

   Remember to remove and retain the Comment Summary Worksheet. You will need this to prepare a final summary evaluation memo (≤ two pages) later in the semester.
5. Return the completed Scorebook to your site coordinator on October 26.

If you have any questions, please call me at (540) 231-XXXX or e-mail at xxxx@yyyy.

Appendix V. Example Cover Memo Distributed for the Control Group's Second Evaluation

To: ____(subject's name)
Course #4

From: Garry D. Coleman, Instructor
Guest Lecturer for Course #4

Re: Colony Fasteners Case Study - Second Evaluation

Date: 26 October 1995

I hope you enjoyed and were challenged by evaluating and scoring the Colony Fasteners Case Study against two of the seven Baldrige criteria categories. You are now asked to evaluate Colony Fasteners against two additional Baldrige criteria categories. Please confirm that you have received a Second Evaluation Scorebook and a questionnaire attached to this memo.

The two additional categories you have been assigned are: ____________. These same category numbers are also listed on the cover of your Second Evaluation Scorebook. *Your evaluation and scores for these two additional categories and the questionnaire are due on November 2, 1995.* A memo (≤ 2 pages) summarizing your evaluation of Colony Fasteners is due on November 9, 1995.

Instructions:

1. Please review the 1995 Baldrige Award Criteria booklet. You should be familiar with the descriptions of the criteria (pp. 2-17); the award examination criteria (pp. 21-39); and the scoring system (pp. 40-41).
2. Read the Colony Fasteners Case Study. Assume Colony Fasteners has asked you to perform an assessment based on the Baldrige criteria. The only information you have about Colony Fasteners is the written case study, which is provided in the form of a mock application for the Baldrige Award.
3. Evaluate the case study against the two new categories you have been assigned using the 1995 Application Scorebook. Please complete the following per the instructions in the Scorebook:
   - the Key Business Factors Worksheet,
   - the applicable Comment and Scoring Worksheets,
   - the Comment Summary Worksheet,
   - the Score Summary Worksheet, and
   - the Conflict of Interest Statement.

   Remember to remove and retain the Comment Summary Worksheet. You will need this to prepare your summary evaluation memo.
4. Complete the questionnaire attached to this memo.
5. Return your completed Second Evaluation Scorebook and questionnaire to your site coordinator on November 2.
6. Your summary evaluation memo (due November 9) should address the following: your overall evaluation of Colony Fasteners against the Baldrige criteria, lessons learned from evaluating and scoring this case study, how this case study fits into the overall course, and recommendations for improving the Baldrige Award training and the case study evaluation.

Some of the other students in Course #4 have no assignment this week. They will be evaluating two additional categories next week. Please do not be concerned about being in one group or the other. Each group will be graded independently and then the grades will be normalized for the class.

If you have any questions, please call me at (540) 231-XXXX or e-mail at xxxx@yyyy

Appendix W. Example Memo Distributed to the Treatment Group While the Control Group Receieved Their Second Evaluation

To: ________________________________(subject's name)
Course #4

From: Garry D. Coleman, Instructor
Guest Lecturer for Course #4

Re: Colony Fasteners Case Study

Date: 26 October 1995

I hope you enjoyed and were challenged by evaluating and scoring the Colony Fasteners Case Study against two of the seven Baldrige criteria categories. You have no assignment for this week. Next week, you will be asked to evaluate and score Colony Fasteners against two additional categories of the Baldrige criteria. You will be provided with a second Scorebook at that time. Please retain your copy of the Colony Fasteners Case Study for next week's assignment.

Some of the other students in Course #4 are evaluating two additional categories this week. Please do not be concerned about being in one group or the other. Each group will be graded independently and then the grades will be normalized for the class.

If you have any questions, please call me at (540) 231-XXXX or e-mail at xxxx@yyyy

Appendix X. Example Vignette Used in FOR Training (adapted from NIST, 1994c)

GREAT NORTHERN CASE STUDY

CATEGORY 6.0: BUSINESS RESULTS

6.3 Supplier Performance Results

We expect the same quality from our suppliers as our customers expect from us. Item 5.4 described the various ways in which we work very closely with suppliers in order to ensure that our expectations are met. One of our largest efforts in the past three years has been to increase the number of supplier partnerships. Figure 6.3-1 demonstrates our positive results in increasing the number of supplier partnerships.

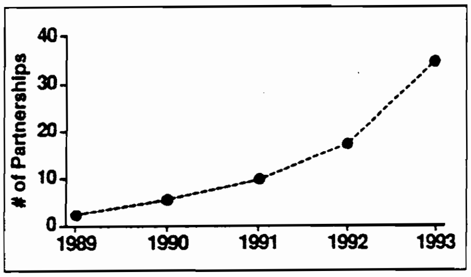


Figure 6.3-1 Number of Supplier Partnerships

The supplier recognition program started in 1991 has been very successful in increasing communication levels between GN and our suppliers. In 1992, 52 suppliers were recognized. Last year, 94 of our suppliers met the high standards defined in our recognition program. This year we hope that all 122 suppliers will attend our recognition celebration.

An important criterion of supplier performance is on-time delivery rates. Although this is measured differently depending on the supplier, we come up with an aggregate measure at the end of each year which tells us how well our suppliers are meeting our on-time delivery standards. Our partnership activities have helped us work with suppliers in order that they may increase their delivery performance. Figure 6.3-2 demonstrates our success over the past few years.

Many of our units have just started to develop and apply quality measures for their group of suppliers. Since these measures are so new, we are not yet able to show trends in these areas. One measure that is representative of these efforts relates to the various independent appraisers that we use across the country. Periodically, our Claims staff goes out to these suppliers and carries out duplicate appraisals to determine the variance between the two quotes. Figure 6.3-3 demonstrates 1993 results for five different appraisal firms: 1) WestApp, 2) Cavelry, 3) Anderson, 4) Gastal, 5) Midstate. The number on the graph is the percent difference between the independent appraiser's quote and GN's staff quote. In most cases, this difference is quite small which reflects on the quality of these suppliers.

While not yet quantified, we know that our supplier partnership activities are doing a great deal to improve the quality performance of our many suppliers.

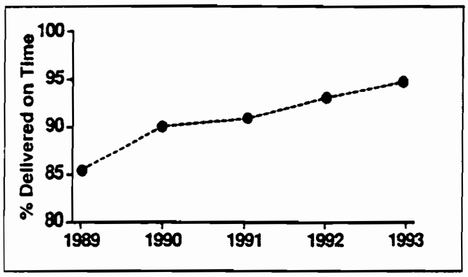


Figure 6.3-2 Supplier On-Time Delivery

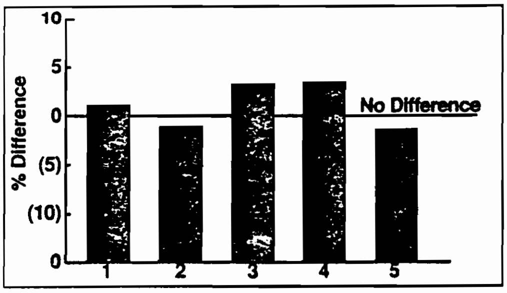


Figure 6.3-3 Independent Appraiser Variance

Appendix Y. Example Memo Distributed to the Control Group While the Treatment Group Received Their Second Evaluation

To: ____(subject's name)
Course #4

From: Garry D. Coleman, Instructor
Guest Lecturer for Course #4

Re: Colony Fasteners Case Study

Date: 2 November 1995

I hope you enjoyed and benefited from evaluating the Colony Fasteners Case Study. The final task of this assignment is to prepare a brief memo (≤ 2 pages) summarizing your evaluation of Colony Fasteners. *This memo is due on November 9.* Your summary evaluation memo should address the following: your overall evaluation of Colony Fasteners against the Baldrige criteria, lessons learned from evaluating and scoring this case study, how this case study fits into the overall course, and recommendations for improving the Baldrige Award training and the case study evaluation.

Some of the other students in Course #4 are evaluating their final two categories this week. Please do not be concerned about being in one group or the other. Each group will be graded independently and then the grades will be normalized for the class.

If you have any questions, please call me at (540) 231-XXXX or e-mail at xxxx@yyyy

Appendix Z. List of Software Packages Used for Data Analyses

Microsoft Excel Version 4.0 for Apple Macintosh. Redmond, WA: Microsoft Corporation.

Microsoft Excel Version 5.0a, Office Professional 7.0 for Windows. Redmond, WA: Microsoft Corporation.

Minitab, Release 10.0 for Windows. State College, PA: Minitab, Inc.

SAS (r) Proprietary Software Release 6.07. Cary, NC: SAS Institute, Inc.

Appendix AA. Edited Minitab Session Files for Hypothesis 1.

Note 1: Because subjects were randomly assigned to the control and treatment groups, equal variances were assumed for the pre-treatment scores.

Note 2: The data used for these analyses were screened per the procedures described under Testing Hypothesis 1. These data are believed to be free from contamination. 1/26/96

```
MTB > TwoSample 95.0 'C1-1.1' 'T1-1.1';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-1.1 vs T1-1.1
           N       Mean      StDev   SE Mean
C1-1.1   12       80.8       18.8       5.4
T1-1.1   13       79.3       16.7       4.6

95% C.I. for mu C1-1.1 - mu T1-1.1: ( -13.2,  16.2)
T-Test mu C1-1.1 = mu T1-1.1 (vs not =): T= 0.21  P=0.83  DF=  23
Both use Pooled StDev = 17.7

MTB > TwoSample 95.0 'C1-1.2' 'T1-1.2';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-1.2 vs T1-1.2
           N       Mean      StDev   SE Mean
C1-1.2   12       77.7       17.3       5.0
T1-1.2   13       80.7       16.9       4.7

95% C.I. for mu C1-1.2 - mu T1-1.2: ( -17.2,  11.1)
T-Test mu C1-1.2 = mu T1-1.2 (vs not =): T= -0.44  P=0.66  DF=  23
Both use Pooled StDev = 17.1

MTB > TwoSample 95.0 'C1-1.3' 'T1-1.3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-1.3 vs T1-1.3
           N       Mean      StDev   SE Mean
C1-1.3   12       72.3       25.2       7.3
T1-1.3   13       81.1       14.4       4.0

95% C.I. for mu C1-1.3 - mu T1-1.3: ( -25.6,  8.1)
T-Test mu C1-1.3 = mu T1-1.3 (vs not =): T= -1.07  P=0.29  DF=  23
Both use Pooled StDev = 20.3

MTB > TwoSample 95.0 'C1-c1.0' 'T1-c1.0';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-c1.0 vs T1-c1.0
           N       Mean      StDev    SE Mean
C1-c1.0   12       76.9       18.3        5.3
T1-c1.0   13       80.4       14.7        4.1

95% C.I. for mu C1-c1.0 - mu T1-c1.0: ( -17.1,  10.2)
T-Test mu C1-c1.0 = mu T1-c1.0 (vs not =): T= -0.52  P=0.61  DF=  23
Both use Pooled StDev = 16.5
```

Descriptive Statistics (1st Evaluation - Category 1.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-1.1           12     80.83     90.00     83.00     18.81      5.43
C1-1.2           12     77.67     80.00     78.20     17.26      4.98
C1-1.3           12     72.33     80.00     73.80     25.22      7.28
C1-c1.0          12     76.94     83.00     78.60     18.26      5.27
T1-1.1           13     79.31     80.00     82.36     16.72      4.64
T1-1.2           13     80.69     85.00     83.55     16.94      4.70
T1-1.3           13     81.08     84.00     82.18     14.44      4.00
T1-c1.0          13     80.36     83.00     83.15     14.65      4.06

Variable        Min       Max        Q1        Q3
C1-1.1        40.00    100.00     72.50     90.00
C1-1.2        50.00    100.00     60.00     91.50
C1-1.3        30.00    100.00     47.50     90.00
C1-c1.0       40.00     97.33     61.67     91.67
T1-1.1        30.00     95.00     75.00     90.00
T1-1.2        30.00    100.00     79.50     90.00
T1-1.3        50.00    100.00     70.00     90.00
T1-c1.0       36.67     93.33     76.67     90.00
```

```
MTB > TwoSample 95.0 'C1-2.1' 'T1-2.1';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-2.1 vs T1-2.1
          N       Mean      StDev    SE Mean
C1-2.1   12       80.0       19.1        5.5
T1-2.1   11       78.3       11.6        3.5

95% C.I. for mu C1-2.1 - mu T1-2.1: ( -12.1,  15.6)
T-Test mu C1-2.1 = mu T1-2.1 (vs not =): T= 0.26  P=0.80  DF=  21
Both use Pooled StDev = 16.0
```

```
MTB > TwoSample 95.0 'C1-2.2' 'T1-2.2';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-2.2 vs T1-2.2
          N       Mean      StDev    SE Mean
C1-2.2   12       78.3       19.5        5.6
```

```
T1-2.2  11       69.8       26.3        7.9
95% C.I. for mu C1-2.2 - mu T1-2.2: ( -11.4,  28.5)
T-Test mu C1-2.2 = mu T1-2.2 (vs not =): T= 0.89  P=0.38  DF=  21
Both use Pooled StDev = 23.0

MTB > TwoSample 95.0 'C1-2.3' 'T1-2.3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-2.3 vs T1-2.3
         N       Mean      StDev   SE Mean
C1-2.3  12       83.3       17.2       5.0
T1-2.3  11       71.0       24.4       7.3

95% C.I. for mu C1-2.3 - mu T1-2.3: ( -5.8,  30.5)
T-Test mu C1-2.3 = mu T1-2.3 (vs not =): T= 1.41  P=0.17  DF=  21
Both use Pooled StDev = 20.9

MTB > TwoSample 95.0 'C1-c2.0' 'T1-c2.0';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-c2.0 vs T1-c2.0
          N       Mean      StDev   SE Mean
C1-c2.0  12       80.6       14.8       4.3
T1-c2.0  11       73.0       16.9       5.1

95% C.I. for mu C1-c2.0 - mu T1-c2.0: ( -6.2,  21.3)
T-Test mu C1-c2.0 = mu T1-c2.0 (vs not =): T= 1.14  P=0.27  DF=  21
Both use Pooled StDev = 15.8

Descriptive Statistics (1st Evaluation - Category 2.0)

Variable          N     Mean   Median   TrMean    StDev   SEMean
C1-2.1           12    80.00    90.00    82.00    19.07     5.50
C1-2.2           12    78.33    80.00    80.00    19.46     5.62
C1-2.3           12    83.33    90.00    85.00    17.23     4.97
C1-c2.0          12    80.56    86.67    81.67    14.83     4.28
T1-2.1           11    78.27    80.00    77.89    11.61     3.50
T1-2.2           11    69.82    80.00    72.00    26.31     7.93
T1-2.3           11    71.00    80.00    72.33    24.35     7.34
T1-c2.0          11    73.03    73.33    72.96    16.90     5.09

Variable        Min      Max       Q1       Q3
C1-2.1        40.00   100.00    65.00    90.00
C1-2.2        40.00   100.00    62.50    97.50
C1-2.3        50.00   100.00    70.00   100.00
C1-c2.0       56.67    93.33    65.83    93.33
T1-2.1        60.00   100.00    70.00    90.00
T1-2.2        20.00   100.00    60.00    90.00
T1-2.3        30.00   100.00    50.00    90.00
T1-c2.0       46.67   100.00    53.33    83.33

MTB > TwoSample 95.0 'C1-3.1' 'T1-3.1';
```

```
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-3.1 vs T1-3.1
          N       Mean      StDev   SE Mean
C1-3.1   11       79.8       16.2       4.9
T1-3.1   13       73.8       18.9       5.3

95% C.I. for mu C1-3.1 - mu T1-3.1: ( -9.1,  21.1)
T-Test mu C1-3.1 = mu T1-3.1 (vs not =): T= 0.82  P=0.42  DF=  22
Both use Pooled StDev = 17.8
```

```
MTB > TwoSample 95.0 'C1-3.2' 'T1-3.2';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-3.2 vs T1-3.2
          N       Mean      StDev   SE Mean
C1-3.2   11       78.2       24.0       7.2
T1-3.2   13       72.3       18.8       5.2

95% C.I. for mu C1-3.2 - mu T1-3.2: ( -12.2,  24.0)
T-Test mu C1-3.2 = mu T1-3.2 (vs not =): T= 0.67  P=0.51  DF=  22
Both use Pooled StDev = 21.3
```

```
MTB > TwoSample 95.0 'C1-c3.0' 'T1-c3.0';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-c3.0 vs T1-c3.0
          N       Mean      StDev   SE Mean
C1-c3.0   11       79.0       18.4       5.6
T1-c3.0   13       73.1       16.8       4.7

95% C.I. for mu C1-c3.0 - mu T1-c3.0: ( -9.0,  20.8)
T-Test mu C1-c3.0 = mu T1-c3.0 (vs not =): T= 0.82  P=0.42  DF=  22
Both use Pooled StDev = 17.5
```

```
MTB > Describe  'C1-3.1'-'T1-c3.0'.
```

Descriptive Statistics (1st Evaluation - Category 3.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-3.1           11     79.82     80.00     80.89     16.23      4.89
C1-3.2           11     78.18     80.00     82.22     24.01      7.24
C1-c3.0          11     79.00     81.50     81.00     18.43      5.56
T1-3.1           13     73.85     80.00     75.45     18.95      5.25
T1-3.2           13     72.31     80.00     72.73     18.78      5.21
T1-c3.0          13     73.08     75.00     73.18     16.78      4.65

Variable        Min       Max        Q1        Q3
C1-3.1        50.00    100.00     70.00     95.00
```

Edited Minitab Session Files for Hypothesis 1 Continued

```
C1-3.2       20.00   100.00    60.00   100.00
C1-c3.0      40.00   100.00    65.00    97.50
T1-3.1       40.00    90.00    60.00    90.00
T1-3.2       40.00   100.00    55.00    90.00
T1-c3.0      50.00    95.00    55.00    90.00
```

```
MTB > TwoSample 95.0 'C1-4.1' 'T1-4.1';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-4.1 vs T1-4.1
         N      Mean     StDev   SE Mean
C1-4.1  11      71.8      17.2       5.2
T1-4.1  14      71.8      21.1       5.6

95% C.I. for mu C1-4.1 - mu T1-4.1: ( -16.2,  16.3)
T-Test mu C1-4.1 = mu T1-4.1 (vs not =): T= 0.00  P=1.0  DF=  23
Both use Pooled StDev = 19.5
```

```
MTB > TwoSample 95.0 'C1-4.2' 'T1-4.2';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-4.2 vs T1-4.2
         N      Mean     StDev   SE Mean
C1-4.2  11      77.3      22.0       6.6
T1-4.2  14      74.3      19.5       5.2

95% C.I. for mu C1-4.2 - mu T1-4.2: ( -14.2,  20.2)
T-Test mu C1-4.2 = mu T1-4.2 (vs not =): T= 0.36  P=0.72  DF=  23
Both use Pooled StDev = 20.6
```

```
MTB > TwoSample 95.0 'C1-4.3' 'T1-4.3';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-4.3 vs T1-4.3
         N      Mean     StDev   SE Mean
C1-4.3  11      83.6      10.3       3.1
T1-4.3  14      75.7      19.5       5.2

95% C.I. for mu C1-4.3 - mu T1-4.3: ( -5.5,  21.4)
T-Test mu C1-4.3 = mu T1-4.3 (vs not =): T= 1.22  P=0.24  DF=  23
Both use Pooled StDev = 16.1
```

```
MTB > TwoSample 95.0 'C1-4.4' 'T1-4.4';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-4.4 vs T1-4.4
         N      Mean     StDev   SE Mean
```

```
C1-4.4  11       81.8       22.3       6.7
T1-4.4  14       77.5       15.8       4.2

95% C.I. for mu C1-4.4 - mu T1-4.4: ( -11.4,  20.1)
T-Test mu C1-4.4 = mu T1-4.4 (vs not =): T= 0.57  P=0.58  DF=  23
Both use Pooled StDev = 18.9
```

```
MTB > TwoSample 95.0 'C1-c4.0' 'T1-c4.0';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-c4.0 vs T1-c4.0
           N       Mean      StDev   SE Mean
C1-c4.0   11       78.6      14.3       4.3
T1-c4.0   14       74.8      17.4       4.7

95% C.I. for mu C1-c4.0 - mu T1-c4.0: ( -9.7,  17.3)
T-Test mu C1-c4.0 = mu T1-c4.0 (vs not =): T= 0.59  P=0.56  DF=  23
Both use Pooled StDev = 16.2
```

Descriptive Statistics (1st Evaluation - Category 4.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-4.1           11     71.82     80.00     71.11     17.22      5.19
C1-4.2           11     77.27     90.00     78.89     21.95      6.62
C1-4.3           11     83.64     80.00     84.44     10.27      3.10
C1-4.4           11     81.82     80.00     86.67     22.28      6.72
C1-c4.0          11     78.64     77.50     80.56     14.33      4.32
T1-4.1           14     71.79     80.00     73.75     21.09      5.64
T1-4.2           14     74.29     80.00     75.00     19.50      5.21
T1-4.3           14     75.71     80.00     77.50     19.50      5.21
T1-4.4           14     77.50     80.00     77.92     15.78      4.22
T1-c4.0          14     74.82     80.00     76.04     17.44      4.66

Variable        Min       Max        Q1        Q3
C1-4.1        50.00    100.00     50.00     80.00
C1-4.2        40.00    100.00     60.00    100.00
C1-4.3        60.00    100.00     80.00     90.00
C1-4.4        20.00    100.00     80.00    100.00
C1-c4.0       42.50     97.50     75.00     87.50
T1-4.1        20.00    100.00     57.50     82.50
T1-4.2        40.00    100.00     50.00     90.00
T1-4.3        30.00    100.00     67.50     90.00
T1-4.4        50.00    100.00     67.50     90.00
T1-c4.0       37.50     97.50     63.13     85.00
```

```
MTB > TwoSample 95.0 'C1-5.1' 'T1-5.1';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-5.1 vs T1-5.1
          N       Mean      StDev   SE Mean
C1-5.1    9       73.3      23.3       7.8
T1-5.1   11       76.8      17.4       5.2
```

Edited Minitab Session Files for Hypothesis 1 Continued

```
95% C.I. for mu C1-5.1 - mu T1-5.1: ( -22.6,  15.6)
T-Test mu C1-5.1 = mu T1-5.1 (vs not =): T= -0.38  P=0.71  DF=  18
Both use Pooled StDev = 20.2

MTB > TwoSample 95.0 'C1-5.2' 'T1-5.2';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-5.2 vs T1-5.2
         N      Mean     StDev   SE Mean
C1-5.2   9      62.2      24.5       8.2
T1-5.2  11      75.0      18.6       5.6

95% C.I. for mu C1-5.2 - mu T1-5.2: ( -33.0,  7.5)
T-Test mu C1-5.2 = mu T1-5.2 (vs not =): T= -1.33  P=0.20  DF=  18
Both use Pooled StDev = 21.4

MTB > TwoSample 95.0 'C1-5.3' 'T1-5.3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-5.3 vs T1-5.3
         N      Mean     StDev   SE Mean
C1-5.3   9      73.9      18.0       6.0
T1-5.3  11      79.1      16.4       4.9

95% C.I. for mu C1-5.3 - mu T1-5.3: ( -21.4,  11.0)
T-Test mu C1-5.3 = mu T1-5.3 (vs not =): T= -0.68  P=0.51  DF=  18
Both use Pooled StDev = 17.1

MTB > TwoSample 95.0 'C1-5.4' 'T1-5.4';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-5.4 vs T1-5.4
         N      Mean     StDev   SE Mean
C1-5.4   9      75.0      16.2       5.4
T1-5.4  11     81.36      8.39       2.5

95% C.I. for mu C1-5.4 - mu T1-5.4: ( -18.2,  5.4)
T-Test mu C1-5.4 = mu T1-5.4 (vs not =): T= -1.13  P=0.27  DF=  18
Both use Pooled StDev = 12.5

MTB > TwoSample 95.0 'C1-c5.0' 'T1-c5.0';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-c5.0 vs T1-c5.0
           N      Mean     StDev   SE Mean
C1-c5.0    9      71.1      16.4       5.5
```

```
T1-c5.0   11      78.1       13.7       4.1

95% C.I. for mu C1-c5.0 - mu T1-c5.0: ( -21.1,  7.1)
T-Test mu C1-c5.0 = mu T1-c5.0 (vs not =): T= -1.04  P=0.31  DF=  18
Both use Pooled StDev = 14.9
```

Descriptive Statistics (1st Evaluation - Category 5.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-5.1            9     73.33     75.00     73.33     23.32      7.77
C1-5.2            9     62.22     70.00     62.22     24.51      8.17
C1-5.3            9     73.89     80.00     73.89     17.99      6.00
C1-5.4            9     75.00     80.00     75.00     16.20      5.40
C1-c5.0           9     71.11     72.50     71.11     16.35      5.45
T1-5.1           11     76.82     80.00     77.22     17.36      5.23
T1-5.2           11     75.00     80.00     77.22     18.57      5.60
T1-5.3           11     79.09     80.00     80.00     16.40      4.95
T1-5.4           11     81.36     80.00     81.67      8.39      2.53
T1-c5.0          11     78.07     80.00     78.75     13.69      4.13

Variable        Min       Max        Q1        Q3
C1-5.1        20.00    100.00     65.00     90.00
C1-5.2        20.00     90.00     37.50     82.50
C1-5.3        35.00    100.00     65.00     80.00
C1-5.4        50.00     90.00     57.50     90.00
C1-c5.0       37.50     95.00     62.50     82.50
T1-5.1        50.00    100.00     60.00     90.00
T1-5.2        30.00    100.00     70.00     90.00
T1-5.3        50.00    100.00     60.00     90.00
T1-5.4        70.00     90.00     70.00     90.00
T1-c5.0       52.50     97.50     72.50     90.00
```

```
MTB > TwoSample 95.0 'C1-6.1' 'T1-6.1';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-6.1 vs T1-6.1
           N       Mean      StDev   SE Mean
C1-6.1     7       83.9       11.1       4.2
T1-6.1    12       84.2       10.8       3.1

95% C.I. for mu C1-6.1 - mu T1-6.1: ( -11.3,  10.7)
T-Test mu C1-6.1 = mu T1-6.1 (vs not =): T= -0.06  P=0.95  DF=  17
Both use Pooled StDev = 10.9
```

```
MTB > TwoSample 95.0 'C1-6.2' 'T1-6.2';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-6.2 vs T1-6.2
           N       Mean      StDev   SE Mean
C1-6.2     7       84.3       14.0       5.3
T1-6.2    12       80.0       12.6       3.6
```

Edited Minitab Session Files for Hypothesis 1 Continued

```
95% C.I. for mu C1-6.2 - mu T1-6.2: ( -8.9,  17.4)
T-Test mu C1-6.2 = mu T1-6.2 (vs not =): T= 0.69  P=0.50  DF=  17
Both use Pooled StDev = 13.1

MTB > TwoSample 95.0 'C1-6.3' 'T1-6.3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-6.3 vs T1-6.3
        N      Mean     StDev   SE Mean
C1-6.3   7      82.9      16.0       6.1
T1-6.3  12      80.8      14.2       4.1

95% C.I. for mu C1-6.3 - mu T1-6.3: ( -12.8,  17.0)
T-Test mu C1-6.3 = mu T1-6.3 (vs not =): T= 0.30  P=0.77  DF=  17
Both use Pooled StDev = 14.9

MTB > TwoSample 95.0 'C1-c6.0' 'T1-c6.0';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-c6.0 vs T1-c6.0
         N      Mean     StDev   SE Mean
C1-c6.0   7     83.67      9.49       3.6
T1-c6.0  12     81.64      9.16       2.6

95% C.I. for mu C1-c6.0 - mu T1-c6.0: ( -7.3,  11.3)
T-Test mu C1-c6.0 = mu T1-c6.0 (vs not =): T= 0.46  P=0.65  DF=  17
Both use Pooled StDev = 9.28

Descriptive Statistics (1st Evaluation - Category 6.0)

Variable          N     Mean   Median   TrMean    StDev   SEMean
C1-6.1            7    83.86    90.00    83.86    11.14     4.21
C1-6.2            7    84.29    90.00    84.29    13.97     5.28
C1-6.3            7    82.86    90.00    82.86    16.04     6.06
C1-c6.0           7    83.67    86.67    83.67     9.49     3.59
T1-6.1           12    84.17    90.00    85.00    10.84     3.13
T1-6.2           12    80.00    80.00    81.00    12.61     3.64
T1-6.3           12    80.75    84.50    82.40    14.22     4.10
T1-c6.0          12    81.64    85.00    81.97     9.16     2.64

Variable        Min      Max       Q1       Q3
C1-6.1        60.00    90.00    80.00    90.00
C1-6.2        60.00   100.00    70.00    90.00
C1-6.3        50.00   100.00    80.00    90.00
C1-c6.0       70.00    95.67    73.33    90.00
T1-6.1        60.00   100.00    80.00    90.00
T1-6.2        50.00   100.00    72.50    88.75
T1-6.3        50.00    95.00    72.50    90.00
T1-c6.0       66.67    93.33    73.33    89.50

MTB > TwoSample 95.0 'C1-7.1' 'T1-7.1';
SUBC>   Alternative 0;
```

Edited Minitab Session Files for Hypothesis 1 Continued

```
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-7.1 vs T1-7.1
         N      Mean     StDev   SE Mean
C1-7.1  12      85.8      16.8       4.8
T1-7.1  12      79.4      13.4       3.9
95% C.I. for mu C1-7.1 - mu T1-7.1: ( -6.4,  19.3)
T-Test mu C1-7.1 = mu T1-7.1 (vs not =): T= 1.04  P=0.31  DF=  22
Both use Pooled StDev = 15.2

MTB > TwoSample 95.0 'C1-7.2' 'T1-7.2';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-7.2 vs T1-7.2
         N      Mean     StDev   SE Mean
C1-7.2  12     88.33      9.37       2.7
T1-7.2  12      84.8      10.5       3.0

95% C.I. for mu C1-7.2 - mu T1-7.2: ( -4.9,  11.9)
T-Test mu C1-7.2 = mu T1-7.2 (vs not =): T= 0.86  P=0.40  DF=  22
Both use Pooled StDev = 9.96

MTB > TwoSample 95.0 'C1-7.3' 'T1-7.3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-7.3 vs T1-7.3
         N      Mean     StDev   SE Mean
C1-7.3  12      75.4      19.0       5.5
T1-7.3  12      68.0      15.4       4.4

95% C.I. for mu C1-7.3 - mu T1-7.3: ( -7.2,  22.1)
T-Test mu C1-7.3 = mu T1-7.3 (vs not =): T= 1.05  P=0.30  DF=  22
Both use Pooled StDev = 17.3

MTB > TwoSample 95.0 'C1-7.4' 'T1-7.4';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-7.4 vs T1-7.4
         N      Mean     StDev   SE Mean
C1-7.4  12      80.4      13.9       4.0
T1-7.4  12      71.7      14.2       4.1

95% C.I. for mu C1-7.4 - mu T1-7.4: ( -3.1,  20.6)
T-Test mu C1-7.4 = mu T1-7.4 (vs not =): T= 1.53  P=0.14  DF=  22
Both use Pooled StDev = 14.0

MTB > TwoSample 95.0 'C1-7.5' 'T1-7.5';
SUBC>   Alternative 0;
```

Edited Minitab Session Files for Hypothesis 1 Continued

```
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-7.5 vs T1-7.5
           N       Mean      StDev   SE Mean
C1-7.5   12      85.83       9.96       2.9
T1-7.5   12       77.3       13.0       3.7

95% C.I. for mu C1-7.5 - mu T1-7.5: ( -1.3,  18.3)
T-Test mu C1-7.5 = mu T1-7.5 (vs not =): T= 1.80  P=0.085  DF=  22
Both use Pooled StDev = 11.6
```

```
MTB > TwoSample 95.0 'C1-c7.0' 'T1-c7.0';
SUBC>   Alternative 0;
SUBC>   Pooled.
```

Two Sample T-Test and Confidence Interval

```
Twosample T for C1-c7.0 vs T1-c7.0
           N       Mean      StDev   SE Mean
C1-c7.0  12      83.17       7.70       2.2
T1-c7.0  12       76.2       10.8       3.1

95% C.I. for mu C1-c7.0 - mu T1-c7.0: ( -1.0,  14.9)
T-Test mu C1-c7.0 = mu T1-c7.0 (vs not =): T= 1.81  P=0.084  DF=  22
Both use Pooled StDev = 9.37
```

Descriptive Statistics (1st Evaluation - Category 7.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.1           12     85.83     90.00     89.00     16.76      4.84
C1-7.2           12     88.33     85.00     88.00      9.37      2.71
C1-7.3           12     75.42     80.00     79.00     19.00      5.49
C1-7.4           12     80.42     80.00     81.50     13.89      4.01
C1-7.5           12     85.83     90.00     86.00      9.96      2.88
C1-c7.0          12     83.17     84.00     84.40      7.70      2.22
T1-7.1           12     79.42     80.00     81.00     13.37      3.86
T1-7.2           12     84.83     90.00     84.80     10.50      3.03
T1-7.3           12     68.00     70.00     67.60     15.40      4.44
T1-7.4           12     71.67     70.00     71.50     14.20      4.10
T1-7.5           12     77.33     80.00     78.80     12.96      3.74
T1-c7.0          12     76.25     76.00     76.90     10.79      3.11

Variable        Min       Max        Q1        Q3
C1-7.1        40.00    100.00     82.50     97.50
C1-7.2        80.00    100.00     80.00    100.00
C1-7.3        25.00     90.00     70.00     90.00
C1-7.4        50.00    100.00     70.00     90.00
C1-7.5        70.00    100.00     80.00     90.00
C1-c7.0       62.00     92.00     80.50     88.00
T1-7.1        50.00     93.00     72.50     90.00
T1-7.2        70.00    100.00     72.50     92.25
T1-7.3        50.00     90.00     50.00     80.00
T1-7.4        50.00     95.00     60.00     83.75
T1-7.5        50.00     90.00     70.00     90.00
T1-c7.0       56.00     90.00     70.00     87.00
```

Appendix AB. Edited Minitab Session Files for Hypothesis 4.

Note: The data used for these analyses were produced by combining the screened control and treatment group scores from Hypothesis 1. Each subject provided scores for two categories in this data set.

Descriptive Statistics (1st Evaluation Combined - Category 1.0)

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
1-1.1          25     80.04     86.00     81.35     17.39      3.48
1-1.2          25     79.24     80.00     80.48     16.81      3.36
1-1.3          25     76.88     80.00     77.91     20.39      4.08
1-c1.0         25     78.72     83.00     79.74     16.23      3.25

Variable      Min       Max        Q1        Q3
1-1.1       30.00    100.00     75.00     90.00
1-1.2       30.00    100.00     70.00     90.00
1-1.3       30.00    100.00     70.00     90.00
1-c1.0      36.67     97.33     75.00     90.00
```

Descriptive Statistics (1st Evaluation Combined - Category 2.0)

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
1-2.1          23     79.17     80.00     80.05     15.62      3.26
1-2.2          23     74.26     80.00     75.62     22.87      4.77
1-2.3          23     77.43     80.00     78.62     21.39      4.46
1-c2.0         23     76.96     83.33     77.30     15.95      3.33

Variable      Min       Max        Q1        Q3
1-2.1       40.00    100.00     70.00     90.00
1-2.2       20.00    100.00     60.00     90.00
1-2.3       30.00    100.00     70.00     91.00
1-c2.0      46.67    100.00     63.33     93.33
```

Descriptive Statistics (1st Evaluation Combined - Category 3.0)

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
1-3.1          24     76.58     80.00     77.18     17.64      3.60
1-3.2          24     75.00     80.00     76.36     21.06      4.30
1-c3.0         24     75.79     77.50     76.32     17.42      3.56

Variable      Min       Max        Q1        Q3
1-3.1       40.00    100.00     62.50     90.00
1-3.2       20.00    100.00     60.00     90.00
1-c3.0      40.00    100.00     65.00     90.00
```

Edited Minitab Session Files for Hypothesis 4 Continued

Descriptive Statistics (1st Evaluation Combined - Category 4.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
1-4.1            25     71.80     80.00     72.83     19.09      3.82
1-4.2            25     75.60     80.00     76.09     20.22      4.04
1.4.3            25     79.20     80.00     80.43     16.31      3.26
1-4.4            25     79.40     80.00     81.09     18.61      3.72
1-c4.0           25     76.50     80.00     77.28     15.94      3.19

Variable        Min       Max        Q1        Q3
1-4.1         20.00    100.00     55.00     80.00
1-4.2         40.00    100.00     55.00     90.00
1.4.3         30.00    100.00     70.00     90.00
1-4.4         20.00    100.00     75.00     90.00
1-c4.0        37.50     97.50     70.00     86.25
```

Descriptive Statistics (1st Evaluation Combined - Category 5.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
1-5.1            20     75.25     80.00     76.94     19.77      4.42
1-5.2            20     69.25     72.50     70.28     21.84      4.88
1-5.3            20     76.75     80.00     77.78     16.88      3.77
1-5.4            20     78.50     80.00     79.44     12.58      2.81
1-c5.0           20     74.94     76.88     75.76     14.96      3.34

Variable        Min       Max        Q1        Q3
1-5.1         20.00    100.00     62.50     90.00
1-5.2         20.00    100.00     62.50     83.75
1-5.3         35.00    100.00     62.50     87.50
1-5.4         50.00     90.00     70.00     90.00
1-c5.0        37.50     97.50     66.25     82.50
```

Descriptive Statistics (1st Evaluation Combined - Category 6.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
1-6.1            19     84.05     90.00     84.53     10.64      2.44
1-6.2            19     81.58     85.00     82.35     12.92      2.96
1-6.3            19     81.53     89.00     82.29     14.50      3.33
1-c6.0           19     82.39     86.67     82.53      9.07      2.08

Variable        Min       Max        Q1        Q3
1-6.1         60.00    100.00     80.00     90.00
1-6.2         50.00    100.00     70.00     90.00
1-6.3         50.00    100.00     80.00     90.00
1-c6.0        66.67     95.67     73.33     90.00
```

Edited Minitab Session Files for Hypothesis 4 Continued

Descriptive Statistics (1st Evaluation Combined - Category 7.0)

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
1-7.1            24     82.62     90.00     83.77     15.19      3.10
1-7.2            24     86.58     90.00     86.73      9.90      2.02
1-7.3            24     71.71     75.00     73.00     17.33      3.54
1-7.4            24     76.04     80.00     76.14     14.44      2.95
1-7.5            24     81.58     81.50     82.18     12.11      2.47
1-c7.0           24     79.71     82.50     80.23      9.82      2.00

Variable        Min       Max        Q1        Q3
1-7.1         40.00    100.00     80.00     90.00
1-7.2         70.00    100.00     80.00     94.50
1-7.3         25.00     90.00     60.00     89.00
1-7.4         50.00    100.00     62.50     90.00
1-7.5         50.00    100.00     71.25     90.00
1-c7.0        56.00     92.00     70.50     88.00

MTB > NoOutfile.
```

Appendix AC. Master and Example Spreadsheets for H4 Pairwise Comparisons

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
1506	1	2	40	50	30	40.0	40	60	70	56.7																							
1555	1	2	80	60	40	60.0	100	40	90	76.7																							
1413	1	2	100	80	80	86.7	90	90	100	93.3																							
2446	1	2	50	80	30	53.3	60	70	60	63.3																							
1448	1	2	90	60	90	80.0	80	70	70	73.3																							
1409	1	2	90	100	90	93.3	90	100	90	93.3																							
1436	1	2	90	100	100	96.7	90	100	90	93.3																							
1425	1	3	90	90	78	86.0					83	80	81.5																				
1513	1	5	80	70	80	76.7													70	70	80	80	75.0										
1528	1	6	90	90	80	86.7																		90	90	80	86.7						
1530	1	6	70	60	70	66.7																		60	70	80	70.0						
1502	1	6	100	92	100	97.3																		87	100	100	95.7						
2508	1	2	80	70	80	76.7	80	90	90	86.7																							
2424	1	2	70	80	70	73.3	80	90	80	83.3																							
2449	1	2	86	79	84	83.0	71	88	91	83.3																							
2411	1	2	90	90	100	93.3	100	100	100	100																							
2537	1	3	80	80	70	76.7					60	70	65.0																				
2505	1	4	90	90	90	90.0								80	90	80	90	85.0															
2531	1	5	90	100	90	93.3													100	100	100	90	97.5										
2545	1	5	95	85	90	90.0													95	90	90	85	90.0										
2546	1	5	90	80	70	80.0													80	80	80	80	80.0										
2536	1	6	70	85	100	85.0																		90	85	95	90.0						
2401	1	6	80	90	70	80.0																		70	80	50	66.7						
2548	1	7	30	30	50	36.7																						70	80	50	50	70	64.0
2515	1	7	80	90	90	86.7																						80	90	90	90	90	88.0
1541	2	3					90	100	80	90.0	100	100	100																				
1442	2	4					90	90	100	93.3				100	100	100	90	97.5															
1552	2	5					50	70	50	56.7									75	85	80	50	72.5										
1431	2	6					90	90	100	93.3														90	60	90	80.0						
1417	2	7					90	60	100	83.3																		100	100	70	70	70	82.0
2437	2	3					90	20	30	46.7	90	100	95.0																				
2414	2	4					90	80	80	83.3				70	90	90	70	80.0															
2504	2	5					70	60	30	53.3									50	60	50	70	57.5										
2419	2	5					60	90	70	73.3									60	80	80	70	72.5										
2433	2	5					70	60	90	73.3									80	70	60	90	75.0										
2507	2	6					70	30	50	50.0														90	70	60	73.3						
2405	2	6					80	60	70	70.0														80	50	80	70.0						
1501	3	4									100	100	100	60	100	90	100	87.5															
1554	3	4									50	80	65.0	50	40	60	20	42.5															
1415	3	4									70	80	75.0	80	70	80	80	77.5															
1410	3	4									80	90	85.0	80	100	80	100	90.0															
1525	3	5									60	20	40.0						20	20	60	50	37.5										
1443	3	5									70	60	65.0						90	80	80	80	82.5										
1441	3	5									80	60	70.0						100	90	100	90	95.0										
1527	3	7									90	90	90.0															100	100	90	80	80	90.0
1522	3	7									95	100	97.5															80	100	90	100	90	92.0
2551	3	4									70	50	60.0	80	90	90	80	85.0															
2416	3	4									60	40	50.0	50	40	50	50	47.5															
2423	3	4									90	50	70.0	70	70	80	90	77.5															
2427	3	4									70	80	75.0	90	90	80	80	85.0															
2429	3	4									90	90	90.0	100	100	90	100	97.5															
2402	3	6									90	90	90.0											80	90	80	83.3						
2439	3	6									90	90	90.0											90	100	90	93.3						
2532	3	7									40	60	50.0															50	70	50	60	50	56.0
2521	3	7									40	60	50.0															90	90	80	70	90	84.0
2553	3	7									90	80	85.0															80	80	70	60	60	70.0
2407	3	7									80	80	80.0															90	100	80	80	90	88.0
1544	4	5												80	90	80	80	82.5	60	70	70	80	70.0										

Master Spreadsheet, H4PAIRS.XLS

Master and Example Spreadsheets for H4 Pairwise Comparisons Continued

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
1510	4	6												80	60	90	80	77.5						90	90	90	90.0						
1543	4	7												50	90	80	80	75.0										70	80	90	80	100	84.0
1447	4	7												70	50	90	100	77.5										100	100	70	70	80	84.0
1403	4	7												50	60	80	90	70.0										90	80	90	50	90	80.0
1422	4	7												90	90	90	80	87.5										90	80	80	90	100	88.0
2434	4	5												60	70	70	60	65.0	50	30	60	70	52.5										
2408	4	5												80	80	70	90	80.0	90	80	80	80	82.5										
2519	4	6												80	80	100	80	85.0						90	80	90	86.7						
2535	4	6												95	90	100	95	95.0						100	80	95	91.7						
2426	4	6												80	50	70	80	70.0						60	90	70	73.3						
2550	4	7												50	50	60	70	57.5										80	70	60	70	80	72.0
2435	4	7												20	50	30	50	37.5										90	70	50	60	80	70.0
1420	5	7																	75	40	35	90	60.0					40	80	25	95	70	62.0
1440	5	7																	90	70	80	90	82.5					90	80	60	80	90	80.0
1430	5	7																	80	35	80	65	65.0					90	90	70	90	90	86.0
2404	5	6																	90	90	100	90	92.5	90	80	90	86.7						
2533	5	7																	70	75	80	90	78.8					80	95	80	85	75	83.0
2444	5	7																	80	70	90	80	80.0					90	90	70	80	70	80.0
1534	6	7																						80	90	50	73.3	90	80	80	70	90	82.0
1428	6	7																						90	90	90	90.0	90	90	90	90	80	88.0
2509	6	7																						90	85	89	88.0	93	93	86	95	83	90.0
2412	6	7																						80	70	80	76.7	60	90	50	60	90	70.0
						n=25				n=23			n=24					n=25					n=20				n=19						n=24
Means			80.0	79.2	76.9	78.7	79.2	74.3	77.4	77.0	76.6	75.0	75.8	71.8	75.6	79.2	79.4	76.5	75.3	69.3	76.8	78.5	74.9	84.1	81.6	81.5	82.4	82.6	86.6	71.7	76.0	81.6	79.7
Std. Dev.			17.39	16.81	20.39	16.23	15.62	22.87	21.39	15.95	17.64	21.06	17.42	19.09	20.22	16.31	18.61	15.94	19.77	21.84	16.88	12.58	14.96	10.64	12.92	14.50	9.07	15.19	9.90	17.33	14.44	12.11	9.82
Variance						263.3				254.5			303.6					254.2					223.8				82.3						96.5

Master Spreadsheet, H4PAIRS.XLS

Master and Example Spreadsheets for H4 Pairwise Comparisons Continued

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	#
1506	1	2	40	50	30	40.0					1
1555	1	2	80	60	40	60.0					2
1413	1	2					90	90	100	93.3	3
2446	1	2					60	70	60	63.3	4
1448	1	2	90	60	90	80.0					5
1409	1	2					90	100	90	93.3	6
1436	1	2	90	100	100	96.7					7
1425	1	3	90	90	78	86.0					
1513	1	5	80	70	80	76.7					
1528	1	6	90	90	80	86.7					
1530	1	6	70	60	70	66.7					
1502	1	6	100	92	100	97.3					
2508	1	2					80	90	90	86.7	8
2424	1	2					80	90	80	83.3	9
2449	1	2					71	88	91	83.3	10
2411	1	2	90	90	100	93.3					11
2537	1	3	80	80	70	76.7					
2505	1	4	90	90	90	90.0					
2531	1	5	90	100	90	93.3					
2545	1	5	95	85	90	90.0					
2546	1	5	90	80	70	80.0					
2536	1	6	70	85	100	85.0					
2401	1	6	80	90	70	80.0					
2548	1	7	30	30	50	36.7					
2515	1	7	80	90	90	86.7					
1541	2	3					90	100	80	90.0	
1442	2	4					90	90	100	93.3	
1552	2	5					50	70	50	56.7	
1431	2	6					90	90	100	93.3	
1417	2	7					90	60	100	83.3	
2437	2	3					90	20	30	46.7	
2414	2	4					90	80	80	83.3	
2504	2	5					70	60	30	53.3	
2419	2	5					60	90	70	73.3	
2433	2	5					70	60	90	73.3	
2507	2	6					70	30	50	50.0	
2405	2	6					80	60	70	70.0	
Means			80.3	78.5	78.3	79.0	78.4	74.3	75.6	76.1	
Std. Dev.			17.83	18.64	20.34	17.23	12.91	22.73	23.10	16.02	
Variance						**296.9**				**256.5**	
						n=19				n=18	

Example Spreadsheet, H4 comparison of category 1 to 2, H4PAIRS.XLS

Appendix AD. Edited Minitab Sessions Files for H4 Pairwise Comparisons

```
Example Minitab Commands
MTB > CDF 1.15750;
SUBC>   F 18 17.

Cumulative Distribution Function

F distribution with 18 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   1.1575         0.6166


F distribution with 22 d.f. in numerator and 23 d.f. in denominator

        x      P( X <= x)
   1.1508         0.6303


F distribution with 24 d.f. in numerator and 23 d.f. in denominator

        x      P( X <= x)
   1.0053         0.5039


F distribution with 22 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   1.0881         0.5646


F distribution with 22 d.f. in numerator and 15 d.f. in denominator

        x      P( X <= x)
   4.0181         0.9961


F distribution with 23 d.f. in numerator and 22 d.f. in denominator

        x      P( X <= x)
   1.9938         0.9447


F distribution with 22 d.f. in numerator and 21 d.f. in denominator

        x      P( X <= x)
   1.1619         0.6331


F distribution with 23 d.f. in numerator and 21 d.f. in denominator

        x      P( X <= x)
   1.0454         0.5384
```

Edited Minitab Sessions Files for H4 Pairwise Comparisons Continued

```
F distribution with 17 d.f. in numerator and 20 d.f. in denominator

        x      P( X <= x)
   1.1010         0.5856

F distribution with 21 d.f. in numerator and 16 d.f. in denominator

        x      P( X <= x)
   3.0473         0.9866


F distribution with 22 d.f. in numerator and 22 d.f. in denominator

        x      P( X <= x)
   2.5298         0.9828


F distribution with 19 d.f. in numerator and 19 d.f. in denominator

        x      P( X <= x)
   1.2897         0.7077


F distribution with 22 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   1.1636         0.6205


F distribution with 22 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   3.5350         0.9948


F distribution with 20 d.f. in numerator and 20 d.f. in denominator

        x      P( X <= x)
   3.8525         0.9980


F distribution with 23 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   1.2181         0.6574


F distribution with 22 d.f. in numerator and 16 d.f. in denominator

        x      P( X <= x)
   3.2942         0.9911
```

Edited Minitab Sessions Files for H4 Pairwise Comparisons Continued

```
F distribution with 22 d.f. in numerator and 20 d.f. in denominator

        x      P( X <= x)
   2.7733         0.9873


F distribution with 19 d.f. in numerator and 17 d.f. in denominator

        x      P( X <= x)
   2.6023         0.9736


F distribution with 16 d.f. in numerator and 21 d.f. in denominator

        x      P( X <= x)
   2.2999         0.9625


F distribution with 21 d.f. in numerator and 16 d.f. in denominator

        x      P( X <= x)
   1.1501         0.6071
```

Appendix AE. Edited Minitab Session Files for Hypothesis 5

Tests of Dimensional Accuracy - Category 1.0

```
MTB > TwoSample 95.0 'C1-DA1' 'T1-DA1';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA1 vs T1-DA1
         N      Mean     StDev   SE Mean
C1-DA1  12      9.85      4.78       1.4
T1-DA1  13     10.26      4.70       1.3
95% C.I. for mu C1-DA1 - mu T1-DA1: ( -4.3,  3.5)
T-Test mu C1-DA1 = mu T1-DA1 (vs not =): T= -0.22  P=0.83  DF=  23
Both use Pooled StDev = 4.74

MTB > Mann-Whitney 95.0 'C1-DA1' 'T1-DA1';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA1     N =  12     Median =       8.500
T1-DA1     N =  13     Median =      10.300
Point estimate for ETA1-ETA2 is      -0.050
95.3 Percent C.I. for ETA1-ETA2 is (-4.099,2.799)
W = 149.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.7442
The test is significant at 0.7427 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Dimensional Accuracy - Category 2.0

```
MTB > TwoSample 95.0 'C1-DA2' 'T1-DA2';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA2 vs T1-DA2
         N      Mean     StDev   SE Mean
C1-DA2  12     11.69      6.85       2.0
T1-DA2  11     12.42      6.52       2.0
95% C.I. for mu C1-DA2 - mu T1-DA2: ( -6.5,  5.1)
T-Test mu C1-DA2 = mu T1-DA2 (vs not =): T= -0.26  P=0.80  DF=  21
Both use Pooled StDev = 6.70

MTB > Mann-Whitney 95.0 'C1-DA2' 'T1-DA2';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA2     N =  12     Median =       10.55
T1-DA2     N =  11     Median =       10.80
Point estimate for ETA1-ETA2 is       -0.25
95.5 Percent C.I. for ETA1-ETA2 is (-7.00,4.00)
W = 139.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.7818
The test is significant at 0.7805 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Dimensional Accuracy - Category 3.0

```
MTB > TwoSample 95.0 'C1-DA3' 'T1-DA3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA3 vs T1-DA3
         N      Mean     StDev   SE Mean
C1-DA3  11      7.82      5.67       1.7
T1-DA3  13      8.08      4.80       1.3
95% C.I. for mu C1-DA3 - mu T1-DA3: ( -4.7,  4.2)
T-Test mu C1-DA3 = mu T1-DA3 (vs not =): T= -0.12  P=0.90  DF=  22
Both use Pooled StDev = 5.21

MTB > Mann-Whitney 95.0 'C1-DA3' 'T1-DA3';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA3      N =  11     Median =       5.000
T1-DA3      N =  13     Median =       5.000
Point estimate for ETA1-ETA2 is        0.000
95.1 Percent C.I. for ETA1-ETA2 is (-5.000,5.003)
W = 132.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.7721
The test is significant at 0.7620 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Dimensional Accuracy - Category 4.0

```
MTB > TwoSample 95.0 'C1-DA4' 'T1-DA4';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA4 vs T1-DA4
         N      Mean     StDev   SE Mean
C1-DA4  11     15.00      7.59       2.3
T1-DA4  14     10.50      4.46       1.2
95% C.I. for mu C1-DA4 - mu T1-DA4: ( -0.5,  9.5)
T-Test mu C1-DA4 = mu T1-DA4 (vs not =): T= 1.85  P=0.077  DF=  23
Both use Pooled StDev = 6.02

MTB > Mann-Whitney 95.0 'C1-DA4' 'T1-DA4';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA4      N =  11     Median =      16.200
T1-DA4      N =  14     Median =      10.550
Point estimate for ETA1-ETA2 is        5.400
95.4 Percent C.I. for ETA1-ETA2 is (-1.598,11.099)
W = 171.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.1253
The test is significant at 0.1247 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Dimensional Accuracy - Category 5.0

```
MTB > TwoSample 95.0 'C1-DA5' 'T1-DA5';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA5 vs T1-DA5
         N      Mean     StDev   SE Mean
C1-DA5   9     18.11      9.37       3.1
T1-DA5  11     14.13      6.09       1.8

95% C.I. for mu C1-DA5 - mu T1-DA5: ( -3.3,  11.3)
T-Test mu C1-DA5 = mu T1-DA5 (vs not =): T= 1.15  P=0.27  DF=  18
Both use Pooled StDev = 7.72

MTB > Mann-Whitney 95.0 'C1-DA5' 'T1-DA5';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA5     N =   9     Median =      17.50
T1-DA5     N =  11     Median =      16.10
Point estimate for ETA1-ETA2 is       2.80
95.2 Percent C.I. for ETA1-ETA2 is (-3.90,11.40)
W = 109.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.2875
The test is significant at 0.2870 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Dimensional Accuracy - Category 6.0

```
MTB > TwoSample 95.0 'C1-DA6' 'T1-DA6';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA6 vs T1-DA6
         N      Mean     StDev   SE Mean
C1-DA6   7      8.99      5.76       2.2
T1-DA6  12      7.53      5.74       1.7

95% C.I. for mu C1-DA6 - mu T1-DA6: ( -4.3,  7.2)
T-Test mu C1-DA6 = mu T1-DA6 (vs not =): T= 0.53  P=0.60  DF=  17
Both use Pooled StDev = 5.75

MTB > Mann-Whitney 95.0 'C1-DA6' 'T1-DA6';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA6     N =   7     Median =       7.40
T1-DA6     N =  12     Median =       7.35
Point estimate for ETA1-ETA2 is       1.70
95.3 Percent C.I. for ETA1-ETA2 is (-4.70,6.10)
W = 77.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.5541
The test is significant at 0.5524 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Dimensional Accuracy - Category 7.0

```
MTB > TwoSample 95.0 'C1-DA7' 'T1-DA7';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-DA7 vs T1-DA7
         N       Mean      StDev   SE Mean
C1-DA7  12      11.03       5.30       1.5
T1-DA7  12       8.40       3.05      0.88

95% C.I. for mu C1-DA7 - mu T1-DA7: ( -1.0,  6.30)
T-Test mu C1-DA7 = mu T1-DA7 (vs not =): T= 1.49  P=0.15  DF=  22
Both use Pooled StDev = 4.33

MTB > Mann-Whitney 95.0 'C1-DA7' 'T1-DA7';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-DA7      N =  12     Median =       9.700
T1-DA7      N =  12     Median =       7.600
Point estimate for ETA1-ETA2 is        2.000
95.4 Percent C.I. for ETA1-ETA2 is (-0.998,5.503)
W = 174.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.1659
The test is significant at 0.1649 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Elevation - Category 1.0

```
MTB > TwoSample 95.0 'C1-E1' 'T1-E1';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E1 vs T1-E1
        N       Mean      StDev   SE Mean
C1-E1  12       22.5       12.6       3.6
T1-E1  13      25.37       6.61       1.8

95% C.I. for mu C1-E1 - mu T1-E1: ( -11.1,  5.4)
T-Test mu C1-E1 = mu T1-E1 (vs not =): T= -0.72  P=0.48  DF=  23
Both use Pooled StDev = 9.94

MTB > Mann-Whitney 95.0 'C1-E1' 'T1-E1';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E1       N =  12     Median =       24.70
T1-E1       N =  13     Median =       24.70
Point estimate for ETA1-ETA2 is        -1.80
95.3 Percent C.I. for ETA1-ETA2 is (-13.30,6.60)
W = 149.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.7442
The test is significant at 0.7428 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Elevation - Category 2.0

```
MTB > TwoSample 95.0 'C1-E2' 'T1-E2';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E2 vs T1-E2
        N      Mean     StDev   SE Mean
C1-E2  12      22.8      13.9       4.0
T1-E2  11      19.2      10.8       3.2

95% C.I. for mu C1-E2 - mu T1-E2: ( -7.3,  14.4)
T-Test mu C1-E2 = mu T1-E2 (vs not =): T= 0.68  P=0.50  DF=  21
Both use Pooled StDev = 12.5

MTB > Mann-Whitney 95.0 'C1-E2' 'T1-E2';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E2      N =  12     Median =      28.35
T1-E2      N =  11     Median =      15.00
Point estimate for ETA1-ETA2 is       6.70
95.5 Percent C.I. for ETA1-ETA2 is (-9.99,20.01)
W = 157.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.4417
The test is significant at 0.4375 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Elevation - Category 3.0

```
MTB > TwoSample 95.0 'C1-E3' 'T1-E3';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E3 vs T1-E3
        N      Mean     StDev   SE Mean
C1-E3  11      39.0      18.4       5.6
T1-E3  13      33.1      16.8       4.7

95% C.I. for mu C1-E3 - mu T1-E3: ( -9.0,  20.8)
T-Test mu C1-E3 = mu T1-E3 (vs not =): T= 0.82  P=0.42  DF=  22
Both use Pooled StDev = 17.5

MTB > Mann-Whitney 95.0 'C1-E3' 'T1-E3';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E3      N =  11     Median =      41.50
T1-E3      N =  13     Median =      35.00
Point estimate for ETA1-ETA2 is       7.50
95.1 Percent C.I. for ETA1-ETA2 is (-10.00,20.00)
W = 153.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.3848
The test is significant at 0.3825 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Elevation - Category 4.0

```
MTB > TwoSample 95.0 'C1-E4' 'T1-E4';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E4 vs T1-E4
        N      Mean      StDev   SE Mean
C1-E4   11     18.80      7.91       2.4
T1-E4   14     18.09      9.27       2.5

95% C.I. for mu C1-E4 - mu T1-E4: ( -6.5,  8.0)
T-Test mu C1-E4 = mu T1-E4 (vs not =): T= 0.20  P=0.84  DF=  23
Both use Pooled StDev = 8.70

MTB > Mann-Whitney 95.0 'C1-E4' 'T1-E4';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E4      N =  11     Median =      18.80
T1-E4      N =  14     Median =      18.80
Point estimate for ETA1-ETA2 is      -0.00
95.4 Percent C.I. for ETA1-ETA2 is (-7.50,7.50)
W = 144.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.9563
The test is significant at 0.9560 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Elevation - Category 5.0

```
MTB > TwoSample 95.0 'C1-E5' 'T1-E5';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E5 vs T1-E5
        N      Mean      StDev   SE Mean
C1-E5    9     19.1      10.4        3.5
T1-E5   11     22.5      12.4        3.7

95% C.I. for mu C1-E5 - mu T1-E5: ( -14.4,  7.5)
T-Test mu C1-E5 = mu T1-E5 (vs not =): T= -0.67  P=0.51  DF=  18
Both use Pooled StDev = 11.6

MTB > Mann-Whitney 95.0 'C1-E5' 'T1-E5';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E5      N =   9     Median =      18.80
T1-E5      N =  11     Median =      23.80
Point estimate for ETA1-ETA2 is      -5.00
95.2 Percent C.I. for ETA1-ETA2 is (-14.99,7.50)
W = 85.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.4941
The test is significant at 0.4923 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Edited Minitab Session Files for Hypothesis 5 Continued

Tests of Elevation - Category 6.0

```
MTB > TwoSample 95.0 'C1-E6' 'T1-E6';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E6 vs T1-E6
        N       Mean      StDev   SE Mean
C1-E6   7      28.67       9.50       3.6
T1-E6  12      26.64       9.16       2.6

95% C.I. for mu C1-E6 - mu T1-E6: ( -7.3,  11.3)
T-Test mu C1-E6 = mu T1-E6 (vs not =): T= 0.46  P=0.65  DF=  17
Both use Pooled StDev = 9.28

MTB > Mann-Whitney 95.0 'C1-E6' 'T1-E6';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E6      N =    7     Median =       31.70
T1-E6      N =   12     Median =       30.00
Point estimate for ETA1-ETA2 is         2.20
95.3 Percent C.I. for ETA1-ETA2 is (-6.70,13.30)
W = 75.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.6726
The test is significant at 0.6708 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Tests of Elevation - Category 7.0

```
MTB > TwoSample 95.0 'C1-E7' 'T1-E7';
SUBC>   Alternative 0;
SUBC>   Pooled.

Two Sample T-Test and Confidence Interval

Twosample T for C1-E7 vs T1-E7
        N       Mean      StDev   SE Mean
C1-E7  12      22.17       7.70       2.2
T1-E7  12      16.08       9.38       2.7

95% C.I. for mu C1-E7 - mu T1-E7: ( -1.2,  13.3)
T-Test mu C1-E7 = mu T1-E7 (vs not =): T= 1.74  P=0.096  DF=  22
Both use Pooled StDev = 8.58

MTB > Mann-Whitney 95.0 'C1-E7' 'T1-E7';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

C1-E7      N =   12     Median =       23.00
T1-E7      N =   12     Median =       15.00
Point estimate for ETA1-ETA2 is         7.50
95.4 Percent C.I. for ETA1-ETA2 is (-2.00,16.00)
W = 174.5
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.1659
The test is significant at 0.1636 (adjusted for ties)
Cannot reject at alpha = 0.05
```

Appendix AF. Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6

The following descriptive statistics include both the first and second evaluation scores for the control and treatment groups.

Descriptive Statistics for Item 1.1

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-1.1           12     80.83     90.00     83.00     18.81      5.43
C2-1.1           11     84.09     90.00     86.11     13.93      4.20
T1-1.1           13     79.31     80.00     82.36     16.72      4.64
T2-1.1           11     76.36     80.00     78.89     25.41      7.66

Variable        Min       Max        Q1        Q3
C1-1.1        40.00    100.00     72.50     90.00
C2-1.1        50.00    100.00     80.00     90.00
T1-1.1        30.00     95.00     75.00     90.00
T2-1.1        30.00    100.00     70.00    100.00
```

Descriptive Statistics for Item 1.2

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-1.2           12     77.67     80.00     78.20     17.26      4.98
C2-1.2           11     83.18     90.00     86.11     17.65      5.32
T1-1.2           13     80.69     85.00     83.55     16.94      4.70
T2-1.2           11     77.27     80.00     78.89     19.02      5.74

Variable        Min       Max        Q1        Q3
C1-1.2        50.00    100.00     60.00     91.50
C2-1.2        40.00    100.00     75.00    100.00
T1-1.2        30.00    100.00     79.50     90.00
T2-1.2        40.00    100.00     60.00     90.00
```

Descriptive Statistics for Item 1.3

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-1.3           12     72.33     80.00     73.80     25.22      7.28
C2-1.3           11     81.82     80.00     81.11     10.79      3.25
T1-1.3           13     81.08     84.00     82.18     14.44      4.00
T2-1.3           11     74.55     80.00     75.56     19.68      5.93

Variable        Min       Max        Q1        Q3
C1-1.3        30.00    100.00     47.50     90.00
C2-1.3        70.00    100.00     70.00     90.00
T1-1.3        50.00    100.00     70.00     90.00
T2-1.3        40.00    100.00     70.00     80.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Category 1.0

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-c1.0        12     76.94     83.00     78.60     18.26      5.27
C2-c1.0        11     83.03     86.67     84.07      9.24      2.79
T1-c1.0        13     80.36     83.00     83.15     14.65      4.06
T2-c1.0        11     76.06     83.33     78.15     19.82      5.98

Variable      Min       Max        Q1        Q3
C1-c1.0     40.00     97.33     61.67     91.67
C2-c1.0     63.33     93.33     76.67     90.00
T1-c1.0     36.67     93.33     76.67     90.00
T2-c1.0     36.67     96.67     66.67     90.00
```

Descriptive Statistics for Item 2.1

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-2.1         12     80.00     90.00     82.00     19.07      5.50
C2-2.1          7     71.43     75.00     71.43     24.62      9.30
T1-2.1         11     78.27     80.00     77.89     11.61      3.50
T2-2.1          9     64.44     60.00     64.44     16.67      5.56

Variable      Min       Max        Q1        Q3
C1-2.1      40.00    100.00     65.00     90.00
C2-2.1      20.00     95.00     70.00     90.00
T1-2.1      60.00    100.00     70.00     90.00
T2-2.1      40.00     90.00     50.00     80.00
```

Descriptive Statistics for Item 2.2

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-2.2         12     78.33     80.00     80.00     19.46      5.62
C2-2.2          7     72.86     80.00     72.86     20.59      7.78
T1-2.2         11     69.82     80.00     72.00     26.31      7.93
T2-2.2          9     55.56     60.00     55.56     15.90      5.30

Variable      Min       Max        Q1        Q3
C1-2.2      40.00    100.00     62.50     97.50
C2-2.2      40.00    100.00     60.00     90.00
T1-2.2      20.00    100.00     60.00     90.00
T2-2.2      40.00     90.00     40.00     60.00
```

Descriptive Statistics for Item 2.3

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-2.3         12     83.33     90.00     85.00     17.23      4.97
C2-2.3          7     64.29     70.00     64.29     25.07      9.48
T1-2.3         11     71.00     80.00     72.33     24.35      7.34
T2-2.3          9     62.22     60.00     62.22     14.81      4.94

Variable      Min       Max        Q1        Q3
C1-2.3      50.00    100.00     70.00    100.00
C2-2.3      40.00    100.00     40.00     90.00
T1-2.3      30.00    100.00     50.00     90.00
T2-2.3      40.00     90.00     50.00     70.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Category 2.0

```
Variable         N      Mean    Median    TrMean     StDev    SEMean
C1-c2.0         12     80 56     86.67     81.67     14.83      4.28
C2-c2.0          7     69.52     68.33     69.52     20.34      7.69
T1-c2.0         11     73.03     73.33     72.96     16.90      5.09
T2-c2.0          9     60.74     56.67     60.74     13.52      4.51

Variable       Min       Max        Q1        Q3
C1-c2.0      56.67     93.33     65.83     93.33
C2-c2.0      33.33     98.33     63.33     86.67
T1-c2.0      46.67    100.00     53.33     83.33
T2-c2.0      43.33     90.00     51.67     66.67
```

Descriptive Statistics for Item 3.1

```
Variable         N      Mean    Median    TrMean     StDev    SEMean
C1-3.1          11     79.82     80.00     80.89     16.23      4.89
C2-3.1          10     73.80     80.00     77.25     25.25      7.99
T1-3.1          13     73.85     80.00     75.45     18.95      5.25
T2-3.1           9     64.44     60.00     64.44     16.67      5.56

Variable       Min       Max        Q1        Q3
C1-3.1       50.00    100.00     70.00     95.00
C2-3.1       20.00    100.00     57.50     92.00
T1-3.1       40.00     90.00     60.00     90.00
T2-3.1       40.00     90.00     50.00     80.00
```

Descriptive Statistics for Item 3.2

```
Variable         N      Mean    Median    TrMean     StDev    SEMean
C1-3.2          11     78.18     80.00     82.22     24.01      7.24
C2-3.2          10     69.50     70.00     73.13     25.87      8.18
T1-3.2          13     72.31     80.00     72.73     18.78      5.21
T2-3.2           9     56.67     50.00     56.67     25.98      8.66

Variable       Min       Max        Q1        Q3
C1-3.2       20.00    100.00     60.00    100.00
C2-3.2       10.00    100.00     60.00     90.00
T1-3.2       40.00    100.00     55.00     90.00
T2-3.2       10.00     90.00     40.00     80.00
```

Descriptive Statistics for Category 3.0

```
Variable         N      Mean    Median    TrMean     StDev    SEMean
C1-c3.0         11     79.00     81.50     81.00     18.43      5.56
C2-c3.0         10     71.65     73.75     75.31     24.37      7.71
T1-c3.0         13     73.08     75.00     73.18     16.78      4.65
T2-c3.0          9     60.56     60.00     60.56     18.62      6.21

Variable       Min       Max        Q1        Q3
C1-c3.0      40.00    100.00     65.00     97.50
C2-c3.0      15.00     99.00     62.50     90.00
T1-c3.0      50.00     95.00     55.00     90.00
T2-c3.0      30.00     85.00     47.50     80.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Item 4.1

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-4.1           11     71.82     80.00     71.11     17.22      5.19
C2-4.1            9     87.78     90.00     87.78      6.67      2.22
T1-4.1           14     71.79     80.00     73.75     21.09      5.64
T2-4.1           10     75.00     80.00     77.50     19.58      6.19

Variable        Min       Max        Q1        Q3
C1-4.1        50.00    100.00     50.00     80.00
C2-4.1        80.00    100.00     80.00     90.00
T1-4.1        20.00    100.00     57.50     82.50
T2-4.1        30.00    100.00     67.50     90.00
```

Descriptive Statistics for Item 4.2

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-4.2           11     77.27     90.00     78.89     21.95      6.62
C2-4.2            9     83.89     85.00     83.89     13.18      4.39
T1-4.2           14     74.29     80.00     75.00     19.50      5.21
T2-4.2           10     71.00     75.00     72.50     18.53      5.86

Variable        Min       Max        Q1        Q3
C1-4.2        40.00    100.00     60.00    100.00
C2-4.2        60.00    100.00     75.00     95.00
T1-4.2        40.00    100.00     50.00     90.00
T2-4.2        30.00    100.00     60.00     80.00
```

Descriptive Statistics for Item 4.3

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-4.3           11     83.64     80.00     84.44     10.27      3.10
C2-4.3            9     89.44     90.00     89.44     11.30      3.77
T1-4.3           14     75.71     80.00     77.50     19.50      5.21
T2-4.3           10     81.00     80.00     81.25     11.97      3.79

Variable        Min       Max        Q1        Q3
C1-4.3        60.00    100.00     80.00     90.00
C2-4.3        70.00    100.00     80.00    100.00
T1-4.3        30.00    100.00     67.50     90.00
T2-4.3        60.00    100.00     70.00     90.00
```

Descriptive Statistics for Item 4.4

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-4.4           11     81.82     80.00     86.67     22.28      6.72
C2-4.4            9     86.11     90.00     86.11      8.58      2.86
T1-4.4           14     77.50     80.00     77.92     15.78      4.22
T2-4.4           10     80.00     80.00     80.00     11.55      3.65

Variable        Min       Max        Q1        Q3
C1-4.4        20.00    100.00     80.00    100.00
C2-4.4        70.00    100.00     80.00     90.00
T1-4.4        50.00    100.00     67.50     90.00
T2-4.4        60.00    100.00     70.00     90.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Category 4.0

```
Variable        N     Mean   Median   TrMean    StDev   SEMean
C1-c4.0        11    78.64    77.50    80.56    14.33     4.32
C2-c4.0         9    86.81    88.75    86.81     6.28     2.09
T1-c4.0        14    74.82    80.00    76.04    17.44     4.66
T2-c4.0        10    76.75    81.25    78.13    13.02     4.12

Variable      Min      Max       Q1       Q3
C1-c4.0     42.50    97.50    75.00    87.50
C2-c4.0     77.50    95.00    81.25    92.50
T1-c4.0     37.50    97.50    63.13    85.00
T2-c4.0     47.50    95.00    70.62    83.12
```

Descriptive Statistics for Item 5.1

```
Variable        N     Mean   Median   TrMean    StDev   SEMean
C1-5.1          9    73.33    75.00    73.33    23.32     7.77
C2-5.1         10    86.00    90.00    91.25    21.71     6.86
T1-5.1         11    76.82    80.00    77.22    17.36     5.23
T2-5.1         12    77.50    75.00    77.00    13.57     3.92

Variable      Min      Max       Q1       Q3
C1-5.1      20.00   100.00    65.00    90.00
C2-5.1      30.00   100.00    85.00   100.00
T1-5.1      50.00   100.00    60.00    90.00
T2-5.1      60.00   100.00    70.00    87.50
```

Descriptive Statistics for Item 5.2

```
Variable        N     Mean   Median   TrMean    StDev   SEMean
C1-5.2          9    62.22    70.00    62.22    24.51     8.17
C2-5.2         10    84.30    90.00    90.38    23.91     7.56
T1-5.2         11    75.00    80.00    77.22    18.57     5.60
T2-5.2         12    70.83    75.00    71.00    19.75     5.70

Variable      Min      Max       Q1       Q3
C1-5.2      20.00    90.00    37.50    82.50
C2-5.2      20.00   100.00    79.50   100.00
T1-5.2      30.00   100.00    70.00    90.00
T2-5.2      40.00   100.00    52.50    87.50
```

Descriptive Statistics for Item 5.3

```
Variable        N     Mean   Median   TrMean    StDev   SEMean
C1-5.3          9    73.89    80.00    73.89    17.99     6.00
C2-5.3         10    85.00    90.00    87.50    16.33     5.16
T1-5.3         11    79.09    80.00    80.00    16.40     4.95
T2-5.3         12    70.00    65.00    69.00    14.77     4.26

Variable      Min      Max       Q1       Q3
C1-5.3      35.00   100.00    65.00    80.00
C2-5.3      50.00   100.00    73.75   100.00
T1-5.3      50.00   100.00    60.00    90.00
T2-5.3      50.00   100.00    60.00    80.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Item 5.4

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-5.4          9     75.00     80.00     75.00     16.20      5.40
C2-5.4         10     88.70     92.50     90.88     13.87      4.39
T1-5.4         11     81.36     80.00     81.67      8.39      2.53
T2-5.4         12     73.33     70.00     74.00     14.35      4.14

Variable      Min       Max        Q1        Q3
C1-5.4      50.00     90.00     57.50     90.00
C2-5.4      60.00    100.00     78.00    100.00
T1-5.4      70.00     90.00     70.00     90.00
T2-5.4      50.00     90.00     60.00     90.00
```

Descriptive Statistics for Category 5.0

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-c5.0         9     71.11     72.50     71.11     16.35      5.45
C2-c5.0        10     86.00     92.50     88.75     16.58      5.24
T1-c5.0        11     78.07     80.00     78.75     13.69      4.13
T2-c5.0        12     72.92     72.50     72.00     13.05      3.77

Variable      Min       Max        Q1        Q3
C1-c5.0     37.50     95.00     62.50     82.50
C2-c5.0     50.00    100.00     75.31     98.13
T1-c5.0     52.50     97.50     72.50     90.00
T2-c5.0     57.50     97.50     60.00     81.25
```

Descriptive Statistics for Item 6.1

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-6.1          7     83.86     90.00     83.86     11.14      4.21
C2-6.1         12     73.75     70.00     73.50     15.24      4.40
T1-6.1         12     84.17     90.00     85.00     10.84      3.13
T2-6.1          9     57.78     60.00     57.78     18.56      6.19

Variable      Min       Max        Q1        Q3
C1-6.1      60.00     90.00     80.00     90.00
C2-6.1      50.00    100.00     60.00     88.75
T1-6.1      60.00    100.00     80.00     90.00
T2-6.1      30.00     90.00     45.00     70.00
```

Descriptive Statistics for Item 6.2

```
Variable        N      Mean    Median    TrMean     StDev    SEMean
C1-6.2          7     84.29     90.00     84.29     13.97      5.28
C2-6.2         12     79.58     80.00     80.50      9.64      2.78
T1-6.2         12     80.00     80.00     81.00     12.61      3.64
T2-6.2          9     57.78     60.00     57.78     13.02      4.34

Variable      Min       Max        Q1        Q3
C1-6.2      60.00    100.00     70.00     90.00
C2-6.2      60.00     90.00     71.25     90.00
T1-6.2      50.00    100.00     72.50     88.75
T2-6.2      30.00     80.00     55.00     60.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Item 6.3

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-6.3            7     82.86     90.00     82.86     16.04      6.06
C2-6.3           12     81.25     80.00     81.50     12.99      3.75
T1-6.3           12     80.75     84.50     82.40     14.22      4.10
T2-6.3            9     62.22     60.00     62.22     19.86      6.62

Variable        Min       Max        Q1        Q3
C1-6.3        50.00    100.00     80.00     90.00
C2-6.3        60.00    100.00     70.00     90.00
T1-6.3        50.00     95.00     72.50     90.00
T2-6.3        30.00     90.00     45.00     80.00
```

Descriptive Statistics for Category 6.0

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-c6.0           7     83.67     86.67     83.67      9.49      3.59
C2-c6.0          12     78.19     78.33     77.83     11.18      3.23
T1-c6.0          12     81.64     85.00     81.97      9.16      2.64
T2-c6.0           9     59.26     60.00     59.26     15.16      5.05

Variable        Min       Max        Q1        Q3
C1-c6.0       70.00     95.67     73.33     90.00
C2-c6.0       66.67     93.33     66.67     89.17
T1-c6.0       66.67     93.33     73.33     89.50
T2-c6.0       36.67     83.33     46.67     70.00
```

Descriptive Statistics for Item 7.1

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.1           12     85.83     90.00     89.00     16.76      4.84
C2-7.1            9     81.11     80.00     81.11     15.37      5.12
T1-7.1           12     79.42     80.00     81.00     13.37      3.86
T2-7.1            9     75.56     80.00     75.56     20.07      6.69

Variable        Min       Max        Q1        Q3
C1-7.1        40.00    100.00     82.50     97.50
C2-7.1        60.00    100.00     65.00     95.00
T1-7.1        50.00     93.00     72.50     90.00
T2-7.1        30.00     90.00     65.00     90.00
```

Descriptive Statistics for Item 7.2

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.2           12     88.33     85.00     88.00      9.37      2.71
C2-7.2            9     85.56     90.00     85.56     11.30      3.77
T1-7.2           12     84.83     90.00     84.80     10.50      3.03
T2-7.2            9     84.44     80.00     84.44      8.82      2.94

Variable        Min       Max        Q1        Q3
C1-7.2        80.00    100.00     80.00    100.00
C2-7.2        70.00    100.00     75.00     95.00
T1-7.2        70.00    100.00     72.50     92.25
T2-7.2        70.00    100.00     80.00     90.00
```

Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6 Continued

Descriptive Statistics for Item 7.3

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.3           12     75.42     80.00     79.00     19.00      5.49
C2-7.3            9     83.33     80.00     83.33     11.18      3.73
T1-7.3           12     68.00     70.00     67.60     15.40      4.44
T2-7.3            9     71.11     70.00     71.11     10.54      3.51

Variable        Min       Max        Q1        Q3
C1-7.3        25.00     90.00     70.00     90.00
C2-7.3        60.00    100.00     80.00     90.00
T1-7.3        50.00     90.00     50.00     80.00
T2-7.3        60.00     90.00     60.00     80.00
```

Descriptive Statistics for Item 7.4

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.4           12     80.42     80.00     81.50     13.89      4.01
C2-7.4            9     80.00     90.00     80.00     18.71      6.24
T1-7.4           12     71.67     70.00     71.50     14.20      4.10
T2-7.4            9     63.33     70.00     63.33     22.91      7.64

Variable        Min       Max        Q1        Q3
C1-7.4        50.00    100.00     70.00     90.00
C2-7.4        40.00    100.00     70.00     90.00
T1-7.4        50.00     95.00     60.00     83.75
T2-7.4        20.00     90.00     50.00     85.00
```

Descriptive Statistics for Item 7.5

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-7.5           12     85.83     90.00     86.00      9.96      2.88
C2-7.5            9     83.33     90.00     83.33     18.71      6.24
T1-7.5           12     77.33     80.00     78.80     12.96      3.74
T2-7.5            9     70.00     70.00     70.00     11.18      3.73

Variable        Min       Max        Q1        Q3
C1-7.5        70.00    100.00     80.00     90.00
C2-7.5        40.00    100.00     75.00     95.00
T1-7.5        50.00     90.00     70.00     90.00
T2-7.5        50.00     90.00     65.00     75.00
```

Descriptive Statistics for Category 7.0

```
Variable          N      Mean    Median    TrMean     StDev    SEMean
C1-c7.0          12     83.17     84.00     84.40      7.70      2.22
C2-c7.0           9     82.67     86.00     82.67     11.53      3.84
T1-c7.0          12     76.25     76.00     76.90     10.79      3.11
T2-c7.0           9     72.89     74.00     72.89     12.09      4.03

Variable        Min       Max        Q1        Q3
C1-c7.0       62.00     92.00     80.50     88.00
C2-c7.0       58.00     98.00     77.00     89.00
T1-c7.0       56.00     90.00     70.00     87.00
T2-c7.0       52.00     90.00     64.00     82.00
```

Appendix AG. Edited Minitab Session Files of 2-Way ANOVA Results foı Hypothesis 6

Category 1.0

```
MTB > GLM  'Scores11' = 'Group1.1'  |  'Time1.1';
SUBC>   Means 'Group1.1' 'Time1.1'  'Group1.1' * 'Time1.1'.
General Linear Model

Factor    Levels Values
Group1.1       2      1      2
Time1.1        2      1      2
```

Analysis of Variance for Scores11 (Item 1.1)

```
Source              DF      Seq SS      Adj SS      Adj MS        F       P
Group1.1             1       230.8       250.3       250.3     0.69   0.411
Time1.1              1         0.1         0.3         0.3     0.00   0.978
Group1.1*Time1.1     1       112.4       112.4       112.4     0.31   0.581
Error               43     15641.9     15641.9       363.8
Total               46     15985.2
```

```
Unusual Observations for Scores11

Obs.  Scores11        Fit  Stdev.Fit  Residual    St.Resid
  1     40.000     80.833      5.506   -40.833     -2.24R
 35     30.000     79.308      5.290   -49.308     -2.69R
 38     30.000     76.364      5.751   -46.364     -2.55R
 39     30.000     76.364      5.751   -46.364     -2.55R

R denotes an obs. with a large st. resid.
```

```
Means for Scores11

Group1.1*Time1.1
       1       1      80.83       5.506
       1       2      84.09       5.751
       2       1      79.31       5.290
       2       2      76.36       5.751
```

```
MTB > GLM  'Scores12' = 'Group1.2'  |   'Time1.2';
SUBC>   Means 'Group1.2' 'Time1.2'  'Group1.2' *  'Time1.2'.
General Linear Model

Factor    Levels Values
Group1.2       2      1      2
Time1.2        2      1      2
```

Analysis of Variance for Scores12 (Item 1.2)

```
Source              DF      Seq SS      Adj SS      Adj MS        F       P
Group1.2             1        16.3        24.3        24.3     0.08   0.782
Time1.2              1        10.9        12.8        12.8     0.04   0.840
Group1.2*Time1.2     1       233.4       233.4       233.4     0.75   0.393
Error               43     13455.3     13455.3       312.9
Total               46     13715.8
```

```
Unusual Observations for Scores12

Obs. Scores12        Fit Stdev.Fit  Residual   St.Resid
 18    40.000     83.182     5.334   -43.182     -2.56R
 35    30.000     80.692     4.906   -50.692     -2.98R
 39    40.000     77.273     5.334   -37.273     -2.21R

R denotes an obs. with a large st. resid.

Means for Scores12

Group1.2*Time1.2
       1       1     77.67      5.106
       1       2     83.18      5.334
       2       1     80.69      4.906
       2       2     77.27      5.334

MTB > GLM  'Scores13' = 'Group1.3'  |     'Time1.3';
SUBC>   Means 'Group1.3' 'Time1.3'  'Group1.3' * 'Time1.3'.
General Linear Model

Factor    Levels Values
Group1.3       2      1      2
Time1.3        2      1      2

Analysis of Variance for Scores13 (Item 1.3)

Source             DF     Seq SS     Adj SS     Adj MS       F      P
Group1.3            1       17.3        6.3        6.3    0.02  0.892
Time1.3             1       20.6       25.5       25.5    0.08  0.785
Group1.3*Time1.3    1      749.9      749.9      749.9    2.22  0.144
Error              43    14536.0    14536.0      338.0
Total              46    15323.7

Unusual Observations for Scores13

Obs. Scores13        Fit Stdev.Fit  Residual   St.Resid
  1    30.000     72.333     5.308   -42.333     -2.40R
  4    30.000     72.333     5.308   -42.333     -2.40R

R denotes an obs. with a large st. resid.

Means for Scores13

Group1.3*Time1.3
       1       1     72.33      5.308
       1       2     81.82      5.544
       2       1     81.08      5.099
       2       2     74.55      5.544

MTB > GLM  'Scores10' = 'Group1.0'  |     'Time1.0';
SUBC>   Means 'Group1.0' 'Time1.0'  'Group1.0' *  'Time1.0'.
General Linear Model


Factor    Levels Values
Group1.0       2      1      2
Time1.0        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores10 (Category 1.0)

```
Source              DF      Seq SS      Adj SS      Adj MS       F        P
Group1.0             1        25.2        36.9        36.9    0.14    0.706
Time1.0              1         7.4         9.3         9.3    0.04    0.850
Group1.0*Time1.0     1       315.2       315.2       315.2    1.23    0.274
Error               43     11029.5     11029.5       256.5
Total               46     11377.4
```

```
Unusual Observations for Scores10

Obs.  Scores10        Fit  Stdev.Fit   Residual    St.Resid
  1    40.0000    76.9444     4.6233   -36.9444     -2.41R
 35    36.6667    80.3590     4.4419   -43.6923     -2.84R
 38    43.3333    76.0606     4.8289   -32.7273     -2.14R
 39    36.6667    76.0606     4.8289   -39.3939     -2.58R

R denotes an obs. with a large st. resid.
```

Means for Scores10

```
Group1.0*Time1.0
      1       1       76.94       4.623
      1       2       83.03       4.829
      2       1       80.36       4.442
      2       2       76.06       4.829
```

Category 2.0

```
MTB > GLM 'Scores21' = 'Group2.1' | 'Time2.1';
SUBC>   Means 'Group2.1'* 'Time2.1'.
General Linear Model

Factor    Levels Values
Group2.1       2      1      2
Time2.1        2      1      2
```

Analysis of Variance for Scores21 (Item 2.1)

```
Source              DF      Seq SS      Adj SS      Adj MS       F        P
Group2.1             1       223.8       177.2       177.2    0.55    0.462
Time2.1              1      1206.8      1171.7      1171.7    3.66    0.064
Group2.1*Time2.1     1        64.5        64.5        64.5    0.20    0.656
Error               35     11206.1     11206.1       320.2
Total               38     12701.2
```

```
Unusual Observations for Scores21

Obs.  Scores21       Fit  Stdev.Fit   Residual    St.Resid
  1     40.000    80.000      5.165    -40.000     -2.33R
 17     20.000    71.429      6.763    -51.429     -3.10R
R denotes an obs. with a large st. resid.
```

Means for Scores21

```
Group2.1*Time2.1        Mean        Stdev
      1       1        80.00        5.165
      1       2        71.43        6.763
      2       1        78.27        5.395
      2       2        64.44        5.964
MTB > GLM  'Scores22' = 'Group2.2' |  'Time2.2';
SUBC>   Means 'Group2.2' *  'Time2.2'.
General Linear Model
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Factor    Levels Values
Group2.2       2      1      2
Time2.2        2      1      2

Analysis of Variance for Scores22 (Item 2.2)

Source              DF      Seq SS     Adj SS     Adj MS       F      P
Group2.2             1      1625.4     1556.5     1556.5    3.48  0.071
Time2.2              1       959.2      909.9      909.9    2.03  0.163
Group2.2*Time2.2     1       180.3      180.3      180.3    0.40  0.530
Error               35     15655.4    15655.4      447.3
Total               38     18420.3

Unusual Observations for Scores22

Obs.  Scores22        Fit  Stdev.Fit   Residual    St.Resid
 24     20.000     69.818      6.377    -49.818     -2.47R
R denotes an obs. with a large st. resia.

Means for Scores22

Group2.2*Time2.2       Mean       Stdev
       1        1      78.33      6.105
       1        2      72.86      7.994
       2        1      69.82      6.377
       2        2      55.56      7.050

MTB > GLM  'Scores23' = 'Group2.3' | 'Time2.3';
SUBC>   Means 'Group2.3' *  'Time2.3'.
General Linear Model

Factor    Levels Values
Group2.3       2      1      2
Time2.3        2      1      2

Analysis of Variance for Scores23 (Item 2.3)

Source              DF      Seq SS     Adj SS     Adj MS       F      P
Group2.3             1       836.5      484.0      484.0    1.15  0.291
Time2.3              1      1739.1     1808.1     1808.1    4.30  0.046
Group2.3*Time2.3     1       246.3      246.3      246.3    0.59  0.449
Error               35     14723.7    14723.7      420.7
Total               38     17545.6

Unusual Observations for Scores23

Obs.  Scores23        Fit  Stdev.Fit   Residual    St.Resid
 24     30.000     71.000      6.184    -41.000     -2.10R
 26     30.000     71.000      6.184    -41.000     -2.10R
R denotes an obs. with a large st. resid.

Means for Scores23

Group2.3*Time2.3       Mean       Stdev
       1        1      83.33      5.921
       1        2      64.29      7.752
       2        1      71.00      6.184
       2        2      62.22      6.837
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
MTB > GLM  'Scores20' = 'Group2.0' |  'Time2.0';
SUBC>   Means 'Group2.0' *  'Time2.0'.
General Linear Model

Factor   Levels Values
Group2.0      2      1     2
Time2.0       2      1     2
```

Analysis of Variance for Scores20 (Category 2.0)

```
Source              DF     Seq SS    Adj SS    Adj MS      F      P
Group2.0             1      787.7     621.1     621.1   2.36  0.134
Time2.0              1     1282.0    1270.1    1270.1   4.82  0.035
Group2.0*Time2.0     1        3.7       3.7       3.7   0.01  0.906
Error               35     9216.5    9216.5     263.3
Total               38    11289.9

Unusual Observations for Scores20

Obs. Scores20        Fit Stdev.Fit  Residual   St.Resid
 17    33.333     69.524     6.133   -36.190     -2.41R
R denotes an obs. with a large st. resid.

Means for Scores20

Group2.0*Time2.0       Mean      Stdev
       1        1     80.56      4.684
       1        2     69.52      6.133
       2        1     73.03      4.893
       2        2     60.74      5.409
```

Category 3.0

```
MTB > GLM 'Scores31' = 'Group3.1' | 'Time3.1';
General Linear Model

Factor   Levels Values
Group3.1      2      1     2
Time3.1       2      1     2
```

Analysis of Variance for Scores31 (Item 3.1)

```
Source              DF     Seq SS    Adj SS    Adj MS      F      P
Group3.1             1      519.3     620.0     620.0   1.62  0.210
Time3.1              1      629.6     627.5     627.5   1.64  0.208
Group3.1*Time3.1     1       30.2      30.2      30.2   0.08  0.780
Error               39    14903.2   14903.2     382.1
Total               42    16082.3

Unusual Observations for Scores31

Obs. Scores31        Fit Stdev.Fit  Residual   St.Resid
 21    20.000     73.800     6.182   -53.800     -2.90R
R denotes an obs. with a large st. resid.

Means for Scores31

Group3.1*Time3.1       Mean      Stdev
       1        1     79.82      5.894
       1        2     73.80      6.182
       2        1     73.85      5.422
       2        2     64.44      6.516
MTB > GLM 'Scores32' = 'Group3.2' | 'Time3.2';
General Linear Model
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Factor    Levels Values
Group3.2       2      1      2
Time3.2        2      1      2
```

Analysis of Variance for Scores32 (Item 3.2)

```
Source                DF      Seq SS      Adj SS      Adj MS        F      P
Group3.2               1       711.6       923.5       923.5     1.68  0.202
Time3.2                1      1568.1      1561.2      1561.2     2.84  0.100
Group3.2*Time3.2       1       127.8       127.8       127.8     0.23  0.632
Error                 39     21416.9     21416.9       549.2
Total                 42     23824.4
```

```
Unusual Observations for Scores32

Obs.  Scores32        Fit  Stdev.Fit   Residual    St.Resid
  7     20.000     78.182      7.066    -58.182      -2.60R
 21     10.000     69.500      7.410    -59.500      -2.68R
 36     10.000     56.667      7.811    -46.667      -2.11R
R denotes an obs. with a large st. resid.
```

```
Means for Scores32

Group3.2*Time3.2       Mean      Stdev
       1       1      78.18      7.066
       1       2      69.50      7.410
       2       1      72.31      6.499
       2       2      56.67      7.811

MTB > GLM  'Scores30' = 'Group3.0' |  'Time3.0';
General Linear Model

Factor    Levels Values
Group3.0       2      1      2
Time3.0        2      1      2
```

Analysis of Variance for Scores30 (Category 3.0)

```
Source                DF      Seq SS      Adj SS      Adj MS        F      P
Group3.0               1       611.7       764.2       764.2     2.00  0.165
Time3.0                1      1046.2      1042.0      1042.0     2.73  0.107
Group3.0*Time3.0       1        70.6        70.6        70.6     0.18  0.670
Error                 39     14891.7     14891.7       381.8
Total                 42     16620.2
```

```
Unusual Observations for Scores30

Obs.  Scores30        Fit  Stdev.Fit   Residual    St.Resid
  7     40.000     79.000      5.892    -39.000      -2.09R
 21     15.000     71.650      6.179    -56.650      -3.06R
R denotes an obs. with a large st. resid.
```

```
Means for Scores30

Group3.0*Time3.0       Mean      Stdev
       1       1      79.00      5.892
       1       2      71.65      6.179
       2       1      73.08      5.420
       2       2      60.56      6.514
```

Category 4.0

```
MTB > GLM  'Scores41' = 'Group4.1' | 'Time4.1';
General Linear Model
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Factor    Levels Values
Group4.1       2      1      2
Time4.1        2      1      2
```

Analysis of Variance for Scores41 (Item 4.1)

```
Source               DF      Seq SS     Adj SS      Adj MS       F       P
Group4.1              1       376.5      439.4       439.4    1.40   0.244
Time4.1               1       886.1      984.4       984.4    3.14   0.084
Group4.1*Time4.1      1       435.0      435.0       435.0    1.39   0.246
Error                40     12549.5    12549.5       313.7
Total                43     14247.2

Unusual Observations for Scores41

Obs.  Scores41        Fit  Stdev.Fit   Residual    St.Resid
 34     20.000     71.786      4.734    -51.786      -3.03R
 42     30.000     75.000      5.601    -45.000      -2.68R
R denotes an obs. with a large st. resid.

Means for Scores41

Group4.1*Time4.1       Mean      Stdev
       1        1      71.82     5.341
       1        2      87.78     5.904
       2        1      71.79     4.734
       2        2      75.00     5.601

MTB > GLM  'Scores42' = 'Group4.2' | 'Time4.2';
General Linear Model

Factor    Levels Values
Group4.2       2      1      2
Time4.2        2      1      2
```

Analysis of Variance for Scores42 (Item 4.2)

```
Source               DF      Seq SS     Adj SS      Adj MS       F       P
Group4.2              1       586.7      674.9       674.9    1.90   0.176
Time4.2               1        17.1       29.7        29.7    0.08   0.774
Group4.2*Time4.2      1       262.5      262.5       262.5    0.74   0.396
Error                40     14239.9    14239.9       356.0
Total                43     15106.2

Unusual Observations for Scores42

Obs.  Scores42        Fit  Stdev.Fit   Residual    St.Resid
  3     40.000     77.273      5.689    -37.273      -2.07R
 42     30.000     71.000      5.967    -41.000      -2.29R
R denotes an obs. with a large st. resid.

Means for Scores42
Group4.2*Time4.2       Mean      Stdev
       1        1      77.27     5.689
       1        2      83.89     6.289
       2        1      74.29     5.043
       2        2      71.00     5.967

MTB > GLM  'Scores43' = 'Group4.3' | 'Time4.3';
General Linear Model

Factor    Levels Values
Group4.3       2      1      2
Time4.3        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores43 (Item 4.3)

```
Source              DF     Seq SS     Adj SS     Adj MS        F       P
Group4.3             1      757.6      717.3      717.3     3.45   0.071
Time4.3              1      329.2      329.6      329.6     1.59   0.215
Group4.3*Time4.3     1        0.7        0.7        0.7     0.00   0.953
Error               40     8309.6     8309.6      207.7
Total               43     9397.2

Unusual Observations for Scores43
Obs.  Scores43        Fit  Stdev.Fit  Residual    St.Resid
 34     30.000     75.714      3.852   -45.714     -3.29R
R denotes an obs. with a large st. resid.

Means for Scores43

Group4.3*Time4.3      Mean      Stdev
       1        1     83.64     4.346
       1        2     89.44     4.804
       2        1     75.71     3.852
       2        2     81.00     4.558

MTB > GLM  'Scores44' = 'Group4.4' |  'Time4.4';
General Linear Model
Factor    Levels Values
Group4.4       2      1      2
Time4.4        2      1      2
```

Analysis of Variance for Scores44 (Item 4.4)

```
Source              DF     Seq SS     Adj SS     Adj MS        F       P
Group4.4             1      295.9      291.3      291.3     1.17   0.287
Time4.4              1      119.1      123.6      123.6     0.49   0.486
Group4.4*Time4.4     1        8.6        8.6        8.6     0.03   0.854
Error               40     9990.0     9990.0      249.8
Total               43    10413.6

Unusual Observations for Scores44
Obs.  Scores44        Fit  Stdev.Fit  Residual    St.Resid
  3     20.000     81.818      4.765   -61.818     -4.10R
R denotes an obs. with a large st. resid.

Means for Scores44

Group4.4*Time4.4      Mean      Stdev
       1        1     81.82     4.765
       1        2     86.11     5.268
       2        1     77.50     4.224
       2        2     80.00     4.998

MTB > GLM  'Scores40' = 'Group4.0' | 'Time4.0';
General Linear Model
Factor    Levels Values
Group4.0       2      1      2
Time4.0        2      1      2
```

Analysis of Variance for Scores40 (Category 4.0)

```
Source              DF     Seq SS     Adj SS     Adj MS        F       P
Group4.0             1      487.9      515.2      515.2     2.62   0.113
Time4.0              1      247.8      273.0      273.0     1.39   0.245
Group4.0*Time4.0     1      104.3      104.3      104.3     0.53   0.470
```

```
Error                  40      7851.9       7851.9         196.3
Total                  43      8691.9

Unusual Observations for Scores40

Obs. Scores40        Fit  Stdev.Fit  Residual    St.Resid
  3   42.5000    78.6364     4.2244  -36.1364     -2.71R
 24   47.5000    74.8214     3.7445  -27.3214     -2.02R
 34   37.5000    74.8214     3.7445  -37.3214     -2.76R
 42   47.5000    76.7500     4.4306  -29.2500     -2.20R
R denotes an obs. with a large st. resid.

Means for Scores40

Group4.0*Time4.0       Mean      Stdev
       1       1      78.64      4.224
       1       2      86.81      4.670
       2       1      74.82      3.745
       2       2      76.75      4.431
```

Category 5.0

```
MTB > GLM  'Scores51' = 'Group5.1' |  'Time5.1';
General Linear Model

Factor   Levels Values
Group5.1      2      1      2
Time5.1       2      1      2
```

Analysis of Variance for Scores51 (Item 5.1)

```
Source                 DF      Seq SS       Adj SS        Adj MS        F        P
Group5.1                1        83.1         65.3          65.3     0.18    0.672
Time5.1                 1       389.9        462.4         462.4     1.29    0.263
Group5.1*Time5.1        1       372.7        372.7         372.7     1.04    0.314
Error                  38     13628.6      13628.6         358.6
Total                  41     14474.4

Unusual Observations for Scores51

Obs. Scores51        Fit  Stdev.Fit  Residual    St.Resid
  3    20.000     73.333      6.313   -53.333     -2.99R
 10    30.000     86.000      5.989   -56.000     -3.12R
R denotes an obs. with a large st. resid.

Means for Scores51
Group5.1*Time5.1       Mean      Stdev
       1       1      73.33      6.313
       1       2      86.00      5.989
       2       1      76.82      5.710
       2       2      77.50      5.467

MTB > GLM  'Scores52' = 'Group5.2' |  'Time5.2';

General Linear Model

Factor   Levels Values
Group5.2      2      1      2
Time5.2       2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Analysis of Variance for Scores52 (Item 5.2)

Source               DF      Seq SS      Adj SS      Adj MS        F       P
Group5.2              1        10.7         1.2         1.2     0.00   0.959
Time5.2               1       621.1       832.5       832.5     1.79   0.189
Group5.2*Time5.2      1      1787.4      1787.4      1787.4     3.84   0.057
Error                38     17691.3     17691.3       465.6
Total                41     20110.6

Unusual Observations for Scores52

Obs.  Scores52        Fit  Stdev.Fit  Residual    St.Resid
  3     20.000     62.222      7.192   -42.222     -2.08R
 10     20.000     84.300      6.823   -64.300     -3.14R
 26     30.000     75.000      6.506   -45.000     -2.19R
R denotes an obs. with a large st. resid.

Means for Scores52

Group5.2*Time5.2       Mean      Stdev
       1       1      62.22      7.192
       1       2      84.30      6.823
       2       1      75.00      6.506
       2       2      70.83      6.229

MTB > GLM  'Scores53' = 'Group5.3' |  'Time5.3';
General Linear Model

Factor    Levels Values
Group5.3       2      1      2
Time5.3        2      1      2

Analysis of Variance for Scores53 (Item 5.3)

Source               DF      Seq SS      Adj SS      Adj MS        F       P
Group5.3              1       302.2       249.1       249.1     0.94   0.339
Time5.3               1         0.0        10.6        10.6     0.04   0.843
Group5.3*Time5.3      1      1059.1      1059.1      1059.1     3.99   0.053
Error                38     10079.8     10079.8       265.3
Total                41     11441.1

Unusual Observations for Scores53

Obs.  Scores53        Fit  Stdev.Fit  Residual    St.Resid
  7     35.000     73.889      5.429   -38.889     -2.53R
 13     50.000     85.000      5.150   -35.000     -2.27R
R denotes an obs. with a large st. resid.

Means for Scores53

Group5.3*Time5.3       Mean      Stdev
       1       1      73.89      5.429
       1       2      85.00      5.150
       2       1      79.09      4.911
       2       2      70.00      4.702

MTB > GLM  'Scores54' = 'Group5.4' | 'Time5.4';
General Linear Model

Factor    Levels Values
Group5.4       2      1      2
Time5.4        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores54 (Item 5.4)

```
Source              DF      Seq SS      Adj SS      Adj MS       F       P
Group5.4             1       263.9       210.3       210.3    1.17   0.285
Time5.4              1        33.8        83.4        83.4    0.47   0.499
Group5.4*Time5.4     1      1225.4      1225.4      1225.4    6.84   0.013
Error               38      6803.3      6803.3       179.0
Total               41      8326.4
```

```
Unusual Observations for Scores54

Obs.  Scores54         Fit  Stdev.Fit   Residual    St.Resid
 13     60.000      88.700      4.231    -28.700     -2.26R
R denotes an obs. with a large st. resid.
```

Means for Scores54

```
Group5.4*Time5.4       Mean      Stdev
       1        1      75.00     4.460
       1        2      88.70     4.231
       2        1      81.36     4.034
       2        2      73.33     3.863

MTB > GLM  'Scores50' = 'Group5.0' |  'Time5.0';
General Linear Model

Factor    Levels Values
Group5.0       2      1      2
Time5.0        2      1      2
```

Analysis of Variance for Scores50 (Category 5.0)

```
Source              DF      Seq SS      Adj SS      Adj MS       F       P
Group5.0             1       132.4        97.4        97.4    0.44   0.510
Time5.0              1       160.2       246.1       246.1    1.12   0.297
Group5.0*Time5.0     1      1042.2      1042.2      1042.2    4.74   0.036
Error               38      8359.2      8359.2       220.0
Total               41      9693.9
```

```
Unusual Observations for Scores50

Obs.  Scores50         Fit  Stdev.Fit   Residual    St.Resid
  3     37.500      71.111      4.944    -33.611     -2.40R
 10     50.000      86.000      4.690    -36.000     -2.56R
R denotes an obs. with a large st. resid.
```

Means for Scores50

```
Group5.0*Time5.0       Mean      Stdev
       1        1      71.11     4.944
       1        2      86.00     4.690
       2        1      78.07     4.472
       2        2      72.92     4.282
```

Category 6.0

```
MTB > GLM  'Scores61' = 'Group6.1' |  'Time6.1';
General Linear Model

Factor    Levels Values
Group6.1       2      1      2
Time6.1        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores61 (Item 6.1)

```
Source              DF     Seq SS     Adj SS     Adj MS       F      P
Group6.1             1      212.6      583.2      583.2    2.86  0.100
Time6.1              1     3402.8     3166.5     3166.5   15.51  0.000
Group6.1*Time6.1     1      630.2      630.2      630.2    3.09  0.087
Error               36     7348.3     7348.3      204.1
Total               39    11593.9
```

```
Unusual Observations for Scores61

Obs.  Scores61       Fit  Stdev.Fit   Residual    St.Resid
 37    90.000     57.778      4.762     32.222       2.39R
 38    30.000     57.778      4.762    -27.778      -2.06R
R denotes an obs. with a large st. resid.
```

```
Means for Scores61

Group6.1*Time6.1      Mean        Stdev
       1       1     83.86        5.400
       1       2     73.75        4.124
       2       1     84.17        4.124
       2       2     57.78        4.762

MTB > GLM  'Scores62' = 'Group6.2' |  'Time6.2';
General Linear Model

Factor    Levels Values
Group6.2       2      1      2
Time6.2        2      1      2
```

Analysis of Variance for Scores62 (Item 6.2)

```
Source              DF     Seq SS     Adj SS     Adj MS       F      P
Group6.2             1     1172.0     1618.4     1618.4   10.99  0.002
Time6.2              1     1907.7     1723.4     1723.4   11.71  0.002
Group6.2*Time6.2     1      729.7      729.7      729.7    4.96  0.032
Error               36     5299.9     5299.9      147.2
Total               39     9109.4
```

```
Unusual Observations for Scores62
Obs.  Scores62       Fit  Stdev.Fit   Residual    St.Resid
  4    60.000     84.286      4.586    -24.286      -2.16R
 23    50.000     80.000      3.503    -30.000      -2.58R
 36    30.000     57.778      4.044    -27.778      -2.43R
R denotes an obs. with a large st. resid.
```

```
Means for Scores62
Group6.2*Time6.2      Mean        Stdev
       1       1     84.29        4.586
       1       2     79.58        3.503
       2       1     80.00        3.503
       2       2     57.78        4.044

MTB > GLM  'Scores63' = 'Group6.3' |  'Time6.3';
General Linear Model

Factor    Levels Values
Group6.3       2      1      2
Time6.3        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores63 (Item 6.3)

```
Source                DF     Seq SS     Adj SS     Adj MS       F       P
Group6.3               1      813.8     1061.9     1061.9    4.35   0.044
Time6.3                1     1096.2      963.8      963.8    3.95   0.054
Group6.3*Time6.3       1      680.7      680.7      680.7    2.79   0.103
Error                 36     8778.9     8778.9      243.9
Total                 39    11369.6

Unusual Observations for Scores63

Obs.  Scores63       Fit  Stdev.Fit   Residual    St.Resid
  6     50.000    82.857      5.902    -32.857     -2.27R
 21     50.000    80.750      4.508    -30.750     -2.06R
 38     30.000    62.222      5.205    -32.222     -2.19R
R denotes an obs. with a large st. resid.
```

```
Means for Scores63

Group6.3*Time6.3      Mean      Stdev
      1        1      82.86     5.902
      1        2      81.25     4.508
      2        1      80.75     4.508
      2        2      62.22     5.205

MTB > GLM  'Scores60' = 'Group6.0' |  'Time6.0';
General Linear Model

Factor    Levels Values
Group6.0       2      1      2
Time6.0        2      1      2
```

Analysis of Variance for Scores60 (Item 6.0)

```
Source                DF     Seq SS     Adj SS     Adj MS       F       P
Group6.0               1      664.7     1044.7     1044.7    8.04   0.007
Time6.0                1     2028.6     1844.2     1844.2   14.19   0.001
Group6.0*Time6.0       1      679.6      679.6      679.6    5.23   0.028
Error                 36     4677.9     4677.9      129.9
Total                 39     8050.8

Unusual Observations for Scores60

Obs.  Scores60       Fit  Stdev.Fit   Residual    St.Resid
 34   83.3333    59.2593     3.7997    24.0741      2.24R
 36   36.6667    59.2593     3.7997   -22.5926     -2.10R
R denotes an obs. with a large st. resid.
```

```
Means for Scores60
Group6.0*Time6.0      Mean      Stdev
      1        1      83.67     4.309
      1        2      78.19     3.291
      2        1      81.64     3.291
      2        2      59.26     3.800
```

Category 7.0

```
MTB > GLM  'Scores71' = 'Group7.1' |  'Time7.1';
General Linear Model

Factor    Levels Values
Group7.1       2      1      2
Time7.1        2      1      2
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

Analysis of Variance for Scores71 (Item 7.1)

```
Source               DF     Seq SS     Adj SS     Adj MS       F      P
Group7.1              1      384.0      368.6      368.6    1.38  0.248
Time7.1               1      189.4      189.4      189.4    0.71  0.405
Group7.1*Time7.1      1        1.9        1.9        1.9    0.01  0.933
Error                38    10167.7    10167.7      267.6
Total                41    10743.1
```

```
Unusual Observations for Scores71

Obs.  Scores71        Fit Stdev.Fit   Residual    St.Resid
  8     40.000     85.833     4.722    -45.833      -2.93R
 39     30.000     75.556     5.453    -45.556      -2.95R
R denotes an obs. with a large st. resid.
```

```
Means for Scores71
Group7.1*Time7.1       Mean       Stdev
       1       1      85.83       4.722
       1       2      81.11       5.453
       2       1      79.42       4.722
       2       2      75.56       5.453
```

```
MTB > GLM  'Scores72' = 'Group7.2' |  'Time7.2';
General Linear Model

Factor    Levels Values
Group7.2       2      1      2
Time7.2        2      1      2
```

Analysis of Variance for Scores72 (Item 7.2)

```
Source               DF     Seq SS     Adj SS     Adj MS       F      P
Group7.2              1       64.4       54.7       54.7    0.54  0.466
Time7.2               1       25.8       25.8       25.8    0.26  0.616
Group7.2*Time7.2      1       14.7       14.7       14.7    0.15  0.705
Error                38     3824.8     3824.8      100.7
Total                41     3929.6
```

```
Means for Scores72
Group7.2*Time7.2       Mean       Stdev
       1       1      88.33       2.896
       1       2      85.56       3.344
       2       1      84.83       2.896
       2       2      84.44       3.344
```

```
MTB > GLM  'Scores73' = 'Group7.3' |  'Time7.3';
General Linear Model

Factor    Levels Values
Group7.3       2      1      2
Time7.3        2      1      2
```

Analysis of Variance for Scores73 (Item 7.3)

```
Source               DF     Seq SS     Adj SS     Adj MS       F      P
Group7.3              1      942.9      991.8      991.8    4.45  0.042
Time7.3               1      312.7      312.7      312.7    1.40  0.244
Group7.3*Time7.3      1       59.4       59.4       59.4    0.27  0.609
Error                38     8469.8     8469.8      222.9
Total                41     9784.8
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Unusual Observations for Scores73

Obs. Scores73       Fit Stdev.Fit  Residual   St.Resid
  8   25.000    75.417     4.310   -50.417     -3.53R
R denotes an obs. with a large st. resid.

Means for Scores73
Group7.3*Time7.3      Mean      Stdev
       1       1     75.42      4.310
       1       2     83.33      4.976
       2       1     68.00      4.310
       2       2     71.11      4.976

MTB > GLM  'Scores74' = 'Group7.4' |  'Time7.4';
General Linear Model

Factor   Levels Values
Group7.4      2      1      2
Time7.4       2      1      2

Analysis of Variance for Scores74 (Item 7.4)

Source                DF       Seq SS       Adj SS       Adj MS         F        P
Group7.4               1       1548.2       1661.2       1661.2      5.57    0.024
Time7.4                1        196.9        196.9        196.9      0.66    0.422
Group7.4*Time7.4       1        161.2        161.2        161.2      0.54    0.467
Error                 38      11339.6      11339.6        298.4
Total                 41      13245.8

Unusual Observations for Scores74

Obs. Scores74       Fit Stdev.Fit  Residual   St.Resid
 15   40.000    80.000     5.758   -40.000     -2.46R
 42   20.000    63.333     5.758   -43.333     -2.66R
R denotes an obs. with a large st. resid.

Means for Scores74
Group7.4*Time7.4      Mean      Stdev
       1       1     80.42      4.987
       1       2     80.00      5.758
       2       1     71.67      4.987
       2       2     63.33      5.758

MTB > GLM  'Scores75' = 'Group7.5' |  'Time7.5';
General Linear Model

Factor   Levels Values
Group7.5      2      1      2
Time7.5       2      1      2

Analysis of Variance for Scores75 (Item 7.5)

Source                DF       Seq SS       Adj SS       Adj MS         F        P
Group7.5               1       1173.4       1225.8       1225.8      6.91    0.012
Time7.5                1        248.6        248.6        248.6      1.40    0.244
Group7.5*Time7.5       1         60.1         60.1         60.1      0.34    0.564
Error                 38       6740.3       6740.3        177.4
Total                 41       8222.5
```

Edited Minitab Session Files of 2-Way ANOVA Results for H6 Continued

```
Unusual Observations for Scores75

Obs.  Scores75        Fit  Stdev.Fit   Residual    St.Resid
 15    40.000      83.333      4.439    -43.333      -3.45R
 24    50.000      77.333      3.845    -27.333      -2.14R
R denotes an obs. with a large st. resid.

Means for Scores75
Group7.5*Time7.5       Mean      Stdev
       1       1      85.83      3.845
       1       2      83.33      4.439
       2       1      77.33      3.845
       2       2      70.00      4.439

MTB > GLM  'Scores70' = 'Group7.0' |  'Time7.0';
General Linear Model

Factor   Levels Values
Group7.0      2      1      2
Time7.0       2      1      2
```

Analysis of Variance for Scores70 (Category 7.0)

```
Source                DF      Seq SS      Adj SS       Adj MS         F      P
Group7.0               1       696.2       716.7        716.7      6.54  0.015
Time7.0                1        38.3        38.3         38.3      0.35  0.558
Group7.0*Time7.0       1        21.0        21.0         21.0      0.19  0.664
Error                 38      4164.8      4164.8        109.6
Total                 41      4920.4

Unusual Observations for Scores70

Obs.  Scores70        Fit  Stdev.Fit   Residual    St.Resid
  8    62.0000    83.1667     3.0221   -21.1667      -2.11R
 15    58.0000    82.6667     3.4897   -24.6667      -2.50R
 24    56.0000    76.2500     3.0221   -20.2500      -2.02R
 39    52.0000    72.8889     3.4897   -20.8889      -2.12R
R denotes an obs. with a large st. resid.

Means for Scores70

Group7.0*Time7.0       Mean      Stdev
       1       1      83.17      3.022
       1       2      82.67      3.490
       2       1      76.25      3.022
       2       2      72.89      3.490
```

Appendix AH. Master and Example Spreadsheets for H9 Pairwise Comparisons

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
1428	1	3	90	90	90	90.0					90	90	90.0																				
1421	1	4	80	70	90	80.0								80	80	100	100	90.0															
1440	1	4	90	100	70	86.7								100	90	90	90	92.5															
1543	1	6	90	80	100	90.0																		80	80	90	83.3						
1552	1	6	95	75	70	80.0																		60	70	70	66.7						
1420	1	6	70	40	80	63.3																		70	60	70	66.7						
1410	1	6	90	100	90	93.3																		90	90	90	90.0						
1442	1	6	90	90	90	90.0																		100	80	100	93.3						
1430	1	6	50	90	80	73.3																		50	80	75	68.3						
1544	1	7	80	80	70	76.7																						80	90	80	90	90	86.0
1431	1	7	100	100	70	90.0																						90	100	100	60	100	90.0
1502	2	3					95	100	100	98.3	98	100	99.0																				
1510	2	3					70	80	40	63.3	50	60	55.0																				
1522	2	4					75	60	70	68.3				80	85	80	85	82.5															
1443	2	4					70	60	70	66.7				80	80	70	80	77.5															
1554	2	5					20	40	40	33.3									30	20	70	80	50.0										
1415	2	5					90	80	90	86.7									100	100	100	100	100.0										
1525	2	7					80	90	40	70.0																		60	70	80	40	40	58.0
1417	3	4									90	60	75.0	90	70	100	70	82.5															
1534	3	5									100	80	90.0						90	90	80	100	90.0										
1506	3	5									70	60	65.0						70	80	50	60	65.0										
1403	3	5									60	85	72.5						100	90	90	90	92.5										
1422	3	6									70	60	65.0											60	70	70	66.7						
1530	3	7									90	90	90.0															80	80	90	90	90	86.0
1555	3	7									20	10	15.0															70	90	80	80	90	82.0
1527	4	5												90	100	100	80	92.5	100	90	90	90	92.5										
1441	4	6												90	100	100	90	95.0						70	90	90	83.3						
1413	4	7												90	90	85	90	88.8										100	90	90	90	70	88.0
1448	4	7												90	60	80	90	80.0										60	70	60	80	90	72.0
1501	5	6																	90	95	95	95	93.8	85	90	85	86.7						
1425	5	6																	90	78	75	72	78.8	70	75	75	73.3						
1436	5	6																	100	100	100	100	100.0	90	90	100	93.3						
1541	5	7																	90	100	100	100	97.5					100	100	90	100	100	98.0
1513	6	7																						60	80	60	66.7	90	80	80	90	80	84.0
			84.1	83.2	81.8	83.0	71.4	72.9	64.3	69.5	73.8	69.5	71.7	87.8	83.9	89.4	86.1	86.8	86.0	84.3	85.0	88.7	86.0	73.8	79.6	81.3	78.2	81.1	85.6	83.3	80.0	83.3	82.7
			13.93	17.65	10.79	9.24	24.62	20.59	25.07	20.34	23.25	25.87	24.37	6.67	13.18	11.30	8.58	6.28	21.71	23.91	16.33	13.87	16.58	15.24	9.64	12.99	11.18	15.37	11.30	11.18	18.71	18.71	11.53
						85.5				413.6			593.9					39.5					274.9				125.0						133.0

Master Spreadsheet, H9CPAIRS.XLS

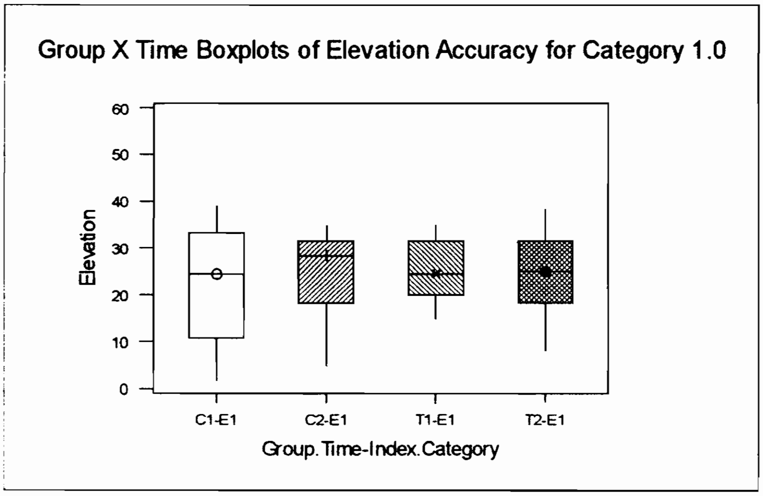
Master and Example Spreadsheets for H9 Pairwise Comparisons Continued

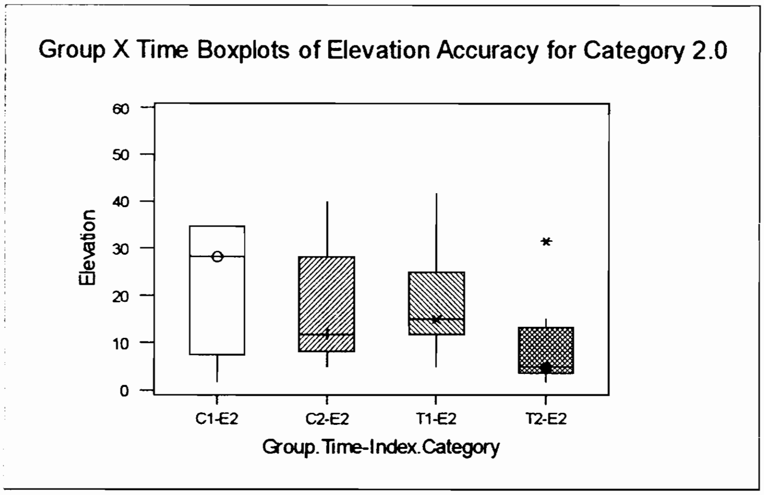
N	Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	#
S	1428	1	3	90	90	90	90.0					
E	1421	1	4	80	70	90	80.0					
H	1440	1	4	90	100	70	86.7					
M	1543	1	6	90	80	100	90.0					
Ri	1552	1	6	95	75	70	80.0					
D	1420	1	6	70	40	80	63.3					
Ri	1410	1	6	90	100	90	93.3					
St	1442	1	6	90	90	90	90.0					
U	1430	1	6	50	90	80	73.3					
N	1544	1	7	80	80	70	76.7					
V	1431	1	7	100	100	70	90.0					
H	1502	2	3					95	100	100	98.3	
Ti	1510	2	3					70	80	40	63.3	
M	1522	2	4					75	60	70	68.3	
A	1443	2	4					70	60	70	66.7	
S	1554	2	5					20	40	40	33.3	
B	1415	2	5					90	80	90	86.7	
B	1525	2	7					80	90	40	70.0	
C	1417	3	4									
Fr	1534	3	5									
G	1506	3	5									
F	1403	3	5									
G	1422	3	6									
C	1530	3	7									
V	1555	3	7									
B	1527	4	5									
Pl	1441	4	6									
A	1413	4	7									
G	1448	4	7									
Fi	1501	5	6									
O	1425	5	6									
P	1436	5	6									
K	1541	5	7									
M	1513	6	7									
Means				84.1	83.2	81.8	83.0	71.4	72.9	64.3	69.5	
Std. Dev.				13.93	17.65	10.79	9.24	24.62	20.59	25.07	20.34	
Variance							**85.5**				**413.6**	
							n=11				n=7	

Example Spreadsheet, H9 comparison of category 1 to 2, H9CPAIRS.XLS

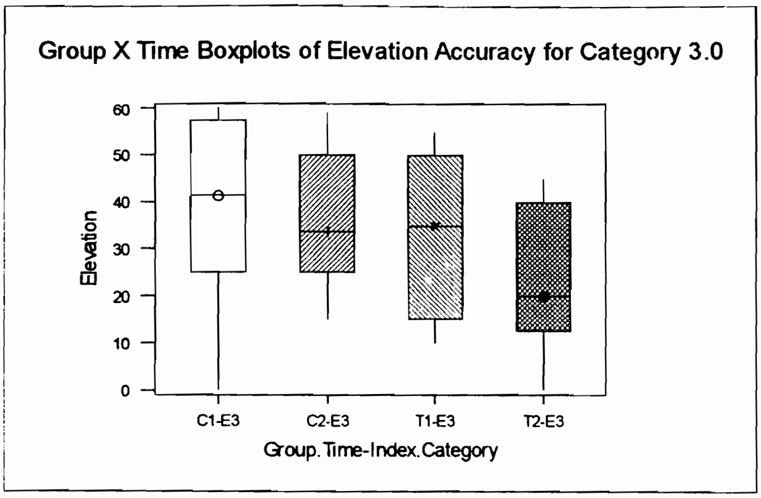
Appendix AI. Group x Time Box Plots of Accuracy Indices by Category

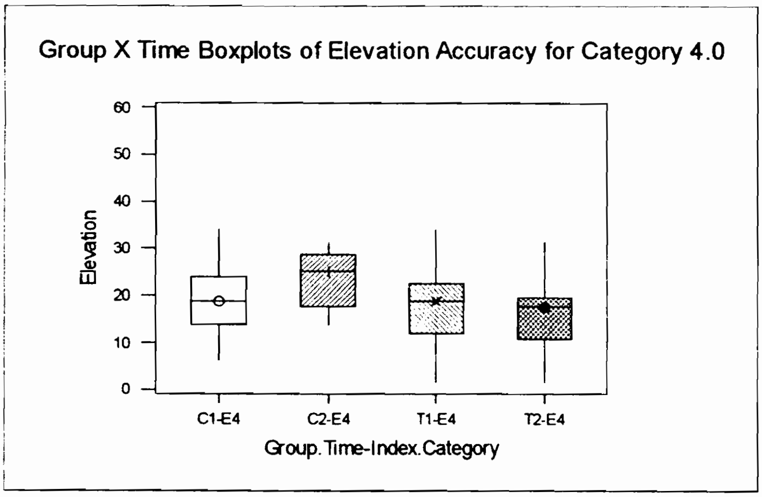
Elevation Box Plots

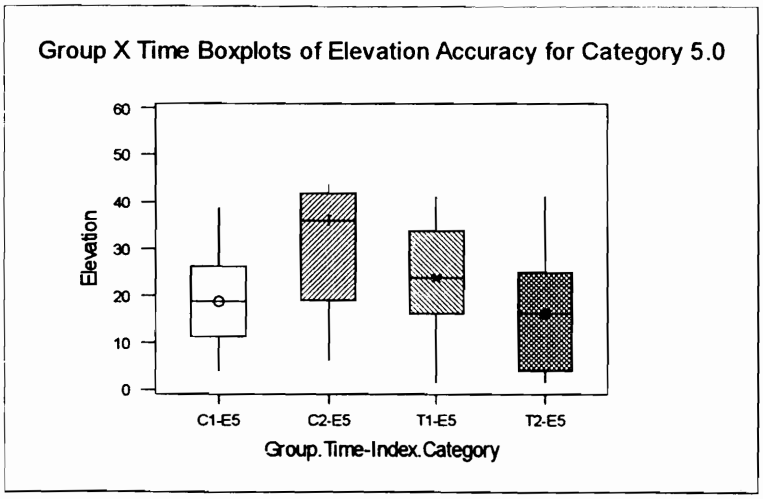

Group X Time Boxplots of Elevation Accuracy for Category 1.0


Group X Time Boxplots of Elevation Accuracy for Category 2.0

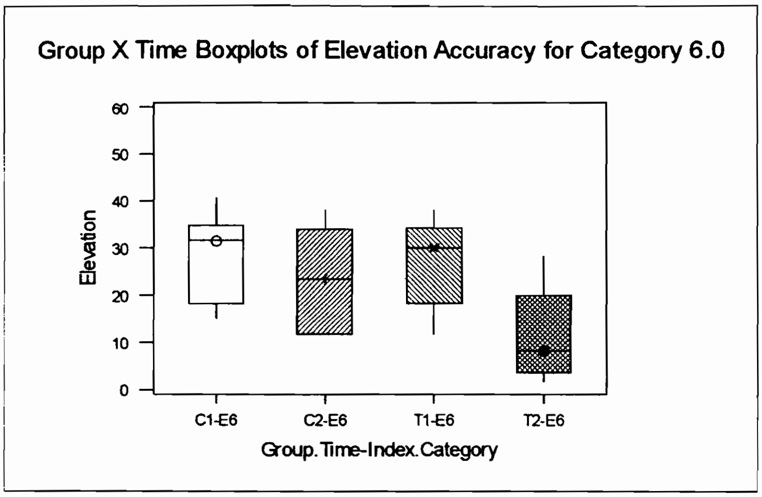
Group x Time Box Plots of Accuracy Indices by Category Continued

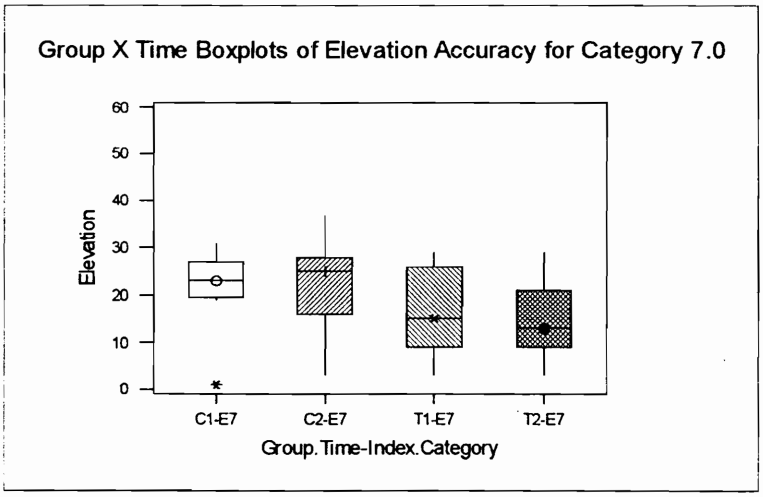






Group x Time Box Plots of Accuracy Indices by Category Continued

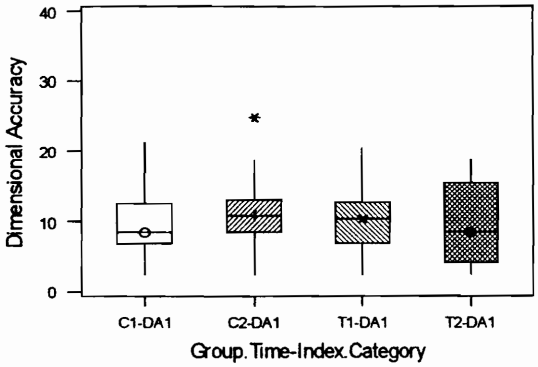




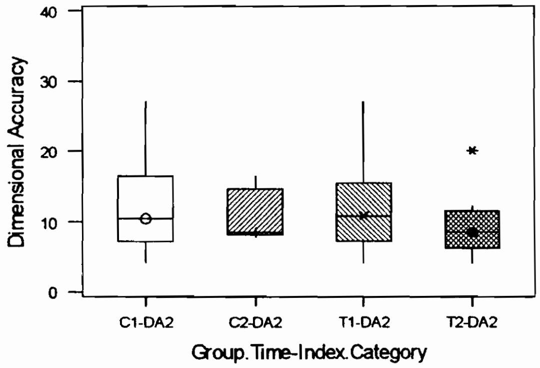
Group x Time Box Plots of Accuracy Indices by Category Continued

Dimensional Accuracy Box Plots

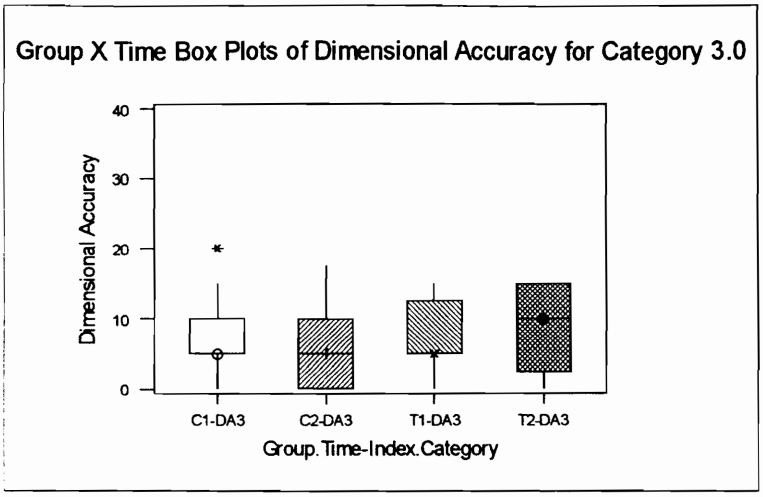
Group X Time Box Plots of Dimensional Accuracy for Category 1.0



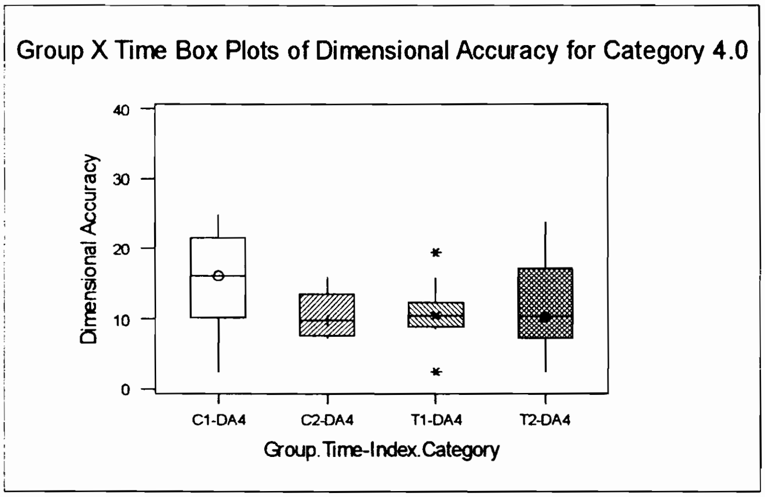
Group X Time Box Plots of Dimensional Accuracy for Category 2.0



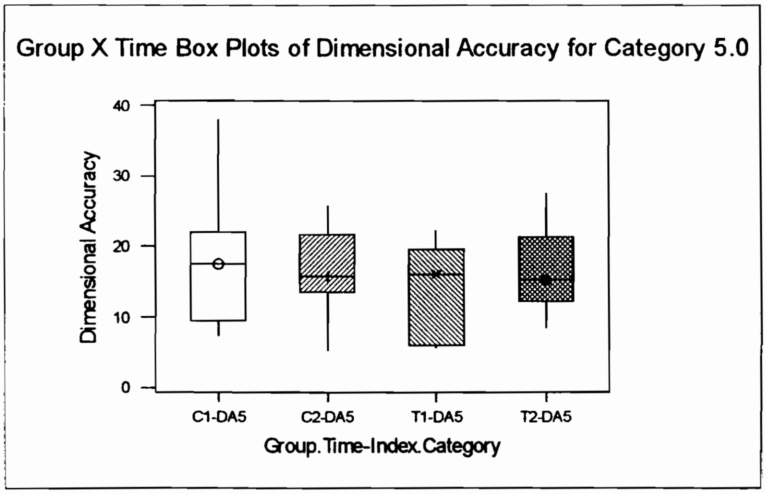
Group x Time Box Plots of Accuracy Indices by Category Continued



Group X Time Box Plots of Dimensional Accuracy for Category 3.0

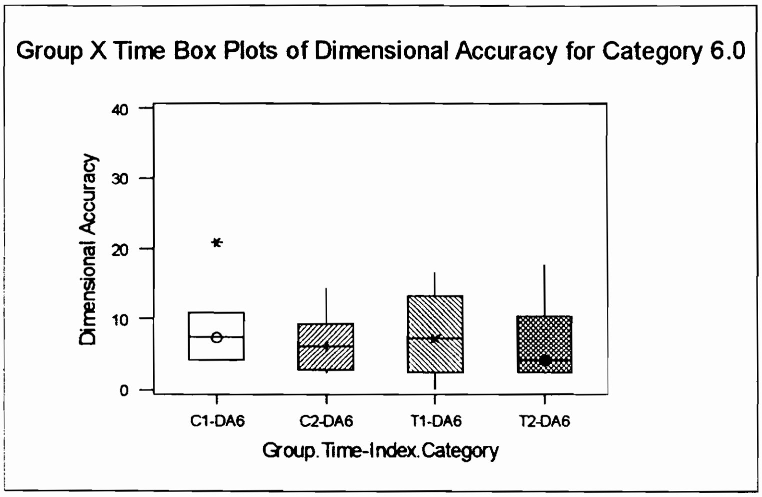


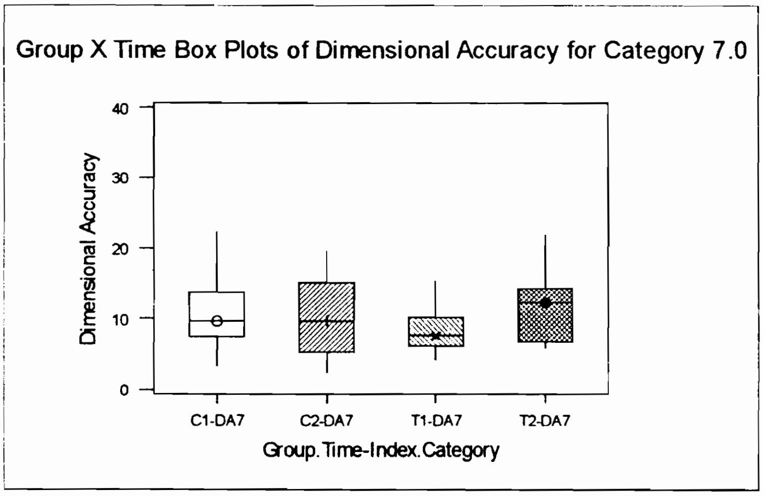
Group X Time Box Plots of Dimensional Accuracy for Category 4.0



Group X Time Box Plots of Dimensional Accuracy for Category 5.0

Group x Time Box Plots of Accuracy Indices by Category Continued





Appendix AJ. Edited Minitab Session Files of 2-Factor ANOVA Results for Hypothesis 10

ANOVAs for Dimensional Accuracy (DA)

```
MTB > RETR 'C:\GARRY\DISSERT\H10DA.MTW'.
Worksheet was saved on  3/21/1996

MTB > GLM 'DAc1.0' = 'Group1.0' | 'Time1.0';
SUBC>   Means 'Group1.0'  'Time1.0'  'Group1.0'*'Time1.0'.

General Linear Model
Factor   Levels Values
Group1.0      2      1     2
Time1.0       2      1     2

Analysis of Variance for DAc1.0  (Dimensional Accuracy - Category 1.0)

Source               DF      Seq SS      Adj SS      Adj MS       F       P
Group1.0              1        3.56        4.19        4.19    0.15   0.701
Time1.0               1        7.03        7.38        7.38    0.26   0.611
Group1.0*Time1.0      1       11.94       11.94       11.94    0.42   0.518
Error                43     1209.64     1209.64       28.13
Total                46     1232.16

Unusual Observations for DAc1.0
Obs.   DAc1.0        Fit Stdev.Fit  Residual   St.Resid
  4   21.2000     9.8500    1.5311   11.3500      2.24R
 21   24.8000    11.6545    1.5992   13.1455      2.60R
 33   20.5000    10.2615    1.4710   10.2385      2.01R
R denotes an obs. with a large st. resid.

Means for DAc1.0
Group1.0              Mean      Stdev
      1             10.752      1.107
      2             10.153      1.086
Time1.0
     1              10.056      1.062
     2              10.850      1.131
Group1.0*Time1.0
      1        1     9.850      1.531
      1        2    11.655      1.599
      2        1    10.262      1.471
      2        2    10.045      1.599

MTB > GLM  'DAc2.0' = 'Group2.0' | 'Time2.0';
SUBC>   Means 'Group2.0' 'Time2.0'   'Group2.0'*'Time2.0'.

General Linear Model
Factor   Levels Values
Group2.0      2      1     2
Time2.0       2      1     2

Analysis of Variance for DAc2.0  (Dimensional Accuracy - Category 2.0)

Source               DF      Seq SS      Adj SS      Adj MS       F       P
Group2.0              1        1.01        1.30        1.30    0.04   0.847
Time2.0               1       33.16       30.91       30.91    0.91   0.348
Group2.0*Time2.0      1       11.28       11.28       11.28    0.33   0.569
Error                35     1194.64     1194.64       34.13
Total                38     1240.09
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

```
Unusual Observations for DAc2.0
Obs.   DAc2.0        Fit Stdev.Fit  Residual   St.Resid
  2   27.2000   11.6917     1.6865   15.5083      2.77R
 24   27.2000   12.4182     1.7615   14.7818      2.65R
R denotes an obs. with a large st. resid.

Means for DAc2.0
Group2.0              Mean      Stdev
       1            11.332      1.389
       2            10.959      1.313
Time2.0
      1             12.055      1.219
      2             10.236      1.472
Group2.0*Time2.0
       1       1    11.692      1.687
       1       2    10.971      2.208
       2       1    12.418      1.762
       2       2     9.500      1.947

MTB > GLM  'DAc3.0' = 'Group3.0' | 'Time3.0';
SUBC>   Means 'Group3.0' 'Time3.0'   'Group3.0'*'Time3.0'.

General Linear Model
Factor   Levels Values
Group3.0      2     1     2
Time3.0       2     1     2

Analysis of Variance for DAc3.0  (Dimensional Accuracy - Category 3.0)

Source              DF      Seq SS      Adj SS      Adj MS       F      P
Group3.0             1       25.09       28.70       28.70    0.92  0.342
Time3.0              1        3.40        3.53        3.53    0.11  0.738
Group3.0*Time3.0     1       20.40       20.40       20.40    0.66  0.423
Error               39     1211.97     1211.97       31.08
Total               42     1260.86

Unusual Observations for DAc3.0

Obs.   DAc3.0        Fit Stdev.Fit  Residual   St.Resid
  4   20.0000    7.8182     1.6808   12.1818      2.29R
 18   17.5000    5.8500     1.7628   11.6500      2.20R
R denotes an obs. with a large st. resid.

Means for DAc3.0
Group3.0              Mean      Stdev
       1             6.834      1.218
       2             8.483      1.209
Time3.0
      1              7.948      1.142
      2              7.369      1.281
Group3.0*Time3.0
       1       1     7.818      1.681
       1       2     5.850      1.763
       2       1     8.077      1.546
       2       2     8.889      1.858

MTB > GLM  'DAc4.0' = 'Group4.0' | 'Time4.0';
SUBC>   Means 'Group4.0' 'Time4.0'   'Group4.0'*'Time4.0'.

General Linear Model
Factor   Levels Values
Group4.0      2     1     2
Time4.0       2     1     2
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

Analysis of Variance for DAc4.0 (Dimensional Accuracy - Category 4.0)

```
Source               DF      Seq SS      Adj SS      Adj MS        F       P
Group4.0              1       47.50       34.11       34.11     1.06   0.310
Time4.0               1       26.69       34.53       34.53     1.07   0.307
Group4.0*Time4.0      1       78.99       78.99       78.99     2.45   0.125
Error                40     1288.72     1288.72       32.22
Total                43     1441.90

Unusual Observations for DAc4.0
Obs.   DAc4.0       Fit Stdev.Fit  Residual    St.Resid
  4    2.4000   15.0000    1.7114  -12.6000      -2.33R
 42   23.8000   11.4200    1.7949   12.3800       2.30R
R denotes an obs. with a large st. resid.

Means for DAc4.0
Group4.0                Mean      Stdev
       1               12.74      1.276
       2               10.96      1.175
Time4.0
      1                12.75      1.143
      2                10.95      1.304
Group4.0*Time4.0
       1       1       15.00      1.711
       1       2       10.49      1.892
       2       1       10.50      1.517
       2       2       11.42      1.795

MTB > GLM  'DAc5.0' = 'Group5.0' | 'Time5.0';
SUBC>   Means 'Group5.0' 'Time5.0'  'Group5.0'*'Time5.0'.

General Linear Model
Factor    Levels Values
Group5.0       2      1      2
Time5.0        2      1      2
```

Analysis of Variance for DAc5.0 (Dimensional Accuracy - Category 5.0)

```
Source               DF      Seq SS      Adj SS      Adj MS        F       P
Group5.0              1       35.18       38.88       38.88     0.82   0.371
Time5.0               1        3.82        1.72        1.72     0.04   0.850
Group5.0*Time5.0      1       43.56       43.56       43.56     0.92   0.343
Error                38     1798.69     1798.69       47.33
Total                41     1881.24

Unusual Observations for DAc5.0
Obs.   DAc5.0       Fit Stdev.Fit  Residual    St.Resid
  2   38.1000   18.1111    2.2933   19.9889       3.08R
R denotes an obs. with a large st. resid.

Means for DAc5.0
Group5.0                Mean      Stdev
       1               17.29      1.581
       2               15.36      1.436
Time5.0
      1                16.12      1.546
      2                16.53      1.473
Group5.0*Time5.0
       1       1       18.11      2.293
       1       2       16.47      2.176
       2       1       14.13      2.074
       2       2       16.58      1.986
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

```
MTB > GLM  'DAc6.0' = 'Group6.0' | 'Time6.0';
SUBC>   Means 'Group6.0' 'Time6.0'  'Group6.0'*'Time6.0'.

General Linear Model
Factor   Levels Values
Group6.0      2     1     2
Time6.0       2     1     2

Analysis of Variance for DAc6.0  (Dimensional Accuracy - Category 6.0)

Source                DF      Seq SS      Adj SS      Adj MS        F       P
Group6.0               1        0.66        3.83        3.83     0.14   0.709
Time6.0                1       23.14       24.89       24.89     0.92   0.344
Group6.0*Time6.0       1        6.49        6.49        6.49     0.24   0.627
Error                 36      973.16      973.16       27.03
Total                 39     1003.45

Unusual Observations for DAc6.0
Obs.    DAc6.0        Fit Stdev.Fit  Residual    St.Resid
  6    20.9000    8.9857    1.9651   11.9143       2.48R
 38    17.8000    6.7333    1.7331   11.0667       2.26R
R denotes an obs. with a large st. resid.

Means for DAc6.0
Group6.0              Mean      Stdev
       1             7.764      1.236
       2             7.129      1.146
Time6.0
      1              8.255      1.236
      2              6.637      1.146
Group6.0*Time6.0
       1       1     8.986      1.965
       1       2     6.542      1.501
       2       1     7.525      1.501
       2       2     6.733      1.733

MTB > GLM  'DAc7.0' = 'Group7.0' | 'Time7.0';
SUBC>   Means 'Group7.0' 'Time7.0'  'Group7.0'*'Time7.0'.

General Linear Model
Factor   Levels Values
Group7.0      2     1     2
Time7.0       2     1     2

Analysis of Variance for DAc7.0  (Dimensional Accuracy - Category 7.0)

Source                DF      Seq SS      Adj SS      Adj MS        F       P
Group7.0               1       10.01        5.04        5.04     0.21   0.648
Time7.0                1       14.27       14.27       14.27     0.60   0.443
Group7.0*Time7.0       1       38.45       38.45       38.45     1.62   0.211
Error                 38      903.75      903.75       23.78
Total                 41      966.47

Unusual Observations for DAc7.0
Obs.    DAc7.0        Fit Stdev.Fit  Residual    St.Resid
  8    22.2000   11.0333    1.4078   11.1667       2.39R
 15    19.6000   10.2778    1.6256    9.3222       2.03R
 42    22.0000   11.5111    1.6256   10.4889       2.28R
R denotes an obs. with a large st. resid.

Means for DAc7.0
Group7.0              Mean      Stdev
       1            10.656     1.0752
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

```
       2            9.956     1.0752
Time7.0
      1             9.717     0.9955
      2            10.894     1.1495
Group7.0*Time7.0
       1       1   11.033     1.4078
       1       2   10.278     1.6256
       2       1    8.400     1.4078
       2       2   11.511     1.6256
```

ANOVAs for Elevation

```
MTB > Retrieve  'C:\GARRY\DISSERT\H10ELEV.MTW'.
Retrieving worksheet from file: C:\GARRY\DISSERT\H10ELEV.MTW
Worksheet was saved on  3/21/1996

MTB > GLM  'Ec1.0' = 'Group1.0' | 'Time1.0';
SUBC>   Means 'Group1.0' 'Time1.0'  'Group1.0'*'Time1.0'.
General Linear Model
Factor   Levels Values
Group1.0      2     1     2
Time1.0       2     1     2

Analysis of Variance for Ec1.0 (Elevation - Category 1.0)

Source               DF      Seq SS      Adj SS      Adj MS        F       P
Group1.0              1       22.00       19.15       19.15     0.21   0.651
Time1.0               1        4.07        4.49        4.49     0.05   0.827
Group1.0*Time1.0      1       29.86       29.86       29.86     0.32   0.573
Error                43     3971.58     3971.58       92.36
Total                46     4027.50

Unusual Observations for Ec1.0
Obs.     Ec1.0        Fit Stdev.Fit  Residual    St.Resid
  2     1.7000    22.4917    2.7743  -20.7917      -2.26R
 18     5.0000    24.7091    2.8977  -19.7091      -2.15R

R denotes an obs. with a large st. resid.

Means for Ec1.0
Group1.0                Mean      Stdev
       1               23.60      2.006
       2               24.88      1.969
Time1.0
      1                23.93      1.924
      2                24.55      2.049
Group1.0*Time1.0
       1       1       22.49      2.774
       1       2       24.71      2.898
       2       1       25.37      2.665
       2       2       24.39      2.898

MTB > GLM  'Ec2.0' = 'Group2.0' | 'Time2.0';
SUBC>   Means 'Group2.0' 'Time2.0'  'Group2.0'*'Time2.0'.

General Linear Model
Factor   Levels Values
Group2.0      2     1     2
Time2.0       2     1     2
```

Analysis of Variance for Ec2.0   (Elevation - Category 2.0)

```
Source              DF      Seq SS      Adj SS      Adj MS        F      P
Group2.0             1       387.0       359.7       359.7     2.52  0.121
Time2.0              1       495.7       473.9       473.9     3.32  0.077
Group2.0*Time2.0     1        66.5        66.5        66.5     0.47  0.499
Error               35      4991.7      4991.7       142.6
Total               38      5940.8

Means for Ec2.0
Group2.0                 Mean      Stdev
       1               20.556      2.840
       2               14.351      2.684
Time2.0
      1                21.014      2.493
      2                13.892      3.009
Group2.0*Time2.0
       1       1       22.783      3.447
       1       2       18.329      4.514
       2       1       19.245      3.601
       2       2        9.456      3.981

MTB > GLM  'Ec3.0' = 'Group3.0' | 'Time3.0';
SUBC>   Means 'Group3.0' 'Time3.0'  'Group3.0'*'Time3.0'.
General Linear Model
Factor   Levels Values
Group3.0      2     1     2
Time3.0       2     1     2
```

Analysis of Variance for Ec3.0   (Elevation - Category 3.0)

```
Source              DF      Seq SS      Adj SS      Adj MS        F      P
Group3.0             1       873.6      1034.1      1034.1     3.80  0.058
Time3.0              1       426.3       422.2       422.2     1.55  0.220
Group3.0*Time3.0     1       166.8       166.8       166.8     0.61  0.438
Error               39     10610.0     10610.0       272.1
Total               42     12076.7

Unusual Observations for Ec3.0
Obs.     Ec3.0         Fit Stdev.Fit   Residual    St.Resid
  7     0.0000    39.0000     4.9731   -39.0000      -2.48R

R denotes an obs. with a large st. resid.

Means for Ec3.0
Group3.0                 Mean      Stdev
       1                37.83      3.603
       2                27.93      3.576
Time3.0
      1                 36.04      3.379
      2                 29.71      3.789
Group3.0*Time3.0
       1       1        39.00      4.973
       1       2        36.65      5.216
       2       1        33.08      4.575
       2       2        22.78      5.498

MTB > GLM  'Ec4.0' = 'Group4.0' | 'Time4.0';
SUBC>   Means 'Group4.0' 'Time4.0'  'Group4.0'*'Time4.0'.
General Linear Model
Factor   Levels Values
Group4.0      2     1     2
Time4.0       2     1     2
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

Analysis of Variance for Ec4.0 (Elevation - Category 4.0)

```
Source              DF      Seq SS      Adj SS      Adj MS        F       P
Group4.0             1      125.62      151.20      151.20     2.30   0.138
Time4.0              1       10.95       16.93       16.93     0.26   0.615
Group4.0*Time4.0     1       99.17       99.17       99.17     1.51   0.227
Error               40     2633.64     2633.64       65.84
Total               43     2869.38

Unusual Observations for Ec4.0
Obs.     Ec4.0        Fit Stdev.Fit  Residual    St.Resid
 27    33.8000    18.0857    2.1686   15.7143       2.01R
 28     1.3000    18.0857    2.1686  -16.7857      -2.15R
R denotes an obs. with a large st. resid.

Means for Ec4.0
Group4.0                Mean      Stdev
       1               20.95      1.824
       2               17.19      1.680
Time4.0
      1                18.44      1.635
      2                19.70      1.864
Group4.0*Time4.0
       1       1       18.80      2.447
       1       2       23.10      2.705
       2       1       18.09      2.169
       2       2       16.30      2.566

MTB > GLM  'Ec5.0' = 'Group5.0' | 'Time5.0';
SUBC>   Means 'Group5.0' 'Time5.0'  'Group5.0'*'Time5.0'.
General Linear Model
Factor   Levels Values
Group5.0      2     1     2
Time5.0       2     1     2
```

Analysis of Variance for Ec5.0 (Elevation - Category 5.0)

```
Source              DF      Seq SS      Adj SS      Adj MS        F       P
Group5.0             1       358.4       305.8       305.8     1.94   0.172
Time5.0              1        51.4        97.6        97.6     0.62   0.436
Group5.0*Time5.0     1       821.4       821.4       821.4     5.20   0.028
Error               38      5996.8      5996.8       157.8
Total               41      7228.0

Unusual Observations for Ec5.0
Obs.     Ec5.0        Fit Stdev.Fit  Residual    St.Resid
 10     6.3000    31.0400    3.9725  -24.7400      -2.08R
 41    41.3000    16.7167    3.6264   24.5833       2.04R
R denotes an obs. with a large st. resid.

Means for Ec5.0
Group5.0                Mean      Stdev
       1               25.06      2.886
       2               19.63      2.622
Time5.0
      1                20.81      2.823
      2                23.88      2.689
Group5.0*Time5.0
       1       1       19.08      4.187
       1       2       31.04      3.973
       2       1       22.55      3.788
       2       2       16.72      3.626
MTB > GLM  'Ec6.0' = 'Group6.0' | 'Time6.0';
```

Edited Minitab Session Files of 2-Way ANOVA Results for Hypothesis 10 Continued

```
SUBC>   Means 'Group6.0' 'Time6.0'  'Group6.0'*'Time6.0'.
General Linear Model
Factor   Levels Values
Group6.0      2      1      2
Time6.0       2      1      2
```

Analysis of Variance for Ec6.0 (Elevation - Category 6.0)

```
Source                DF     Seq SS      Adj SS      Adj MS        F       P
Group6.0               1     232.66      413.48      413.48     4.21   0.047
Time6.0                1    1032.15      959.34      959.34     9.77   0.003
Group6.0*Time6.0       1     198.11      198.11      198.11     2.02   0.164
Error                 36    3533.15     3533.15       98.14
Total                 39    4996.08

Means for Ec6.0
Group6.0             Mean       Stdev
       1             25.93      2.356
       2             19.34      2.184
Time6.0
      1              27.66      2.356
      2              17.61      2.184
Group6.0*Time6.0
       1        1    28.67      3.744
       1        2    23.19      2.860
       2        1    26.64      2.860
       2        2    12.03      3.302

MTB > GLM  'Ec7.0' = 'Group7.0' | 'Time7.0';
SUBC>   Means 'Group7.0' 'Time7.0'  'Group7.0'*'Time7.0'.
General Linear Model
Factor   Levels Values
Group7.0      2      1      2
Time7.0       2      1      2
```

Analysis of Variance for Ec7.0 (Elevation - Category 7.0)

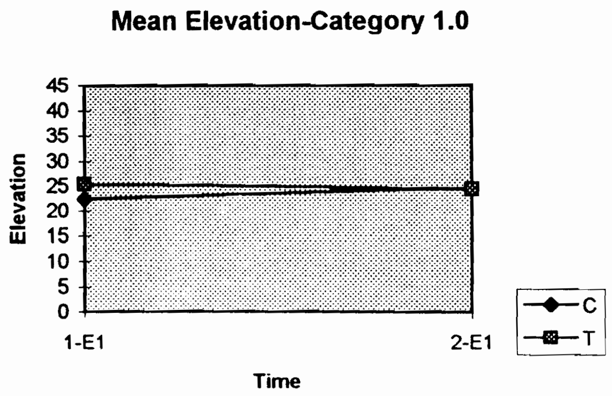
```
Source                DF     Seq SS      Adj SS      Adj MS        F       P
Group7.0               1     486.88      494.05      494.05     6.36   0.016
Time7.0                1       4.76        4.76        4.76     0.06   0.806
Group7.0*Time7.0       1       7.38        7.38        7.38     0.10   0.760
Error                 38    2952.81     2952.81       77.71
Total                 41    3451.83

Unusual Observations for Ec7.0
Obs.     Ec7.0        Fit Stdev.Fit  Residual    St.Resid
  8     1.0000    22.1667    2.5447  -21.1667      -2.51R
 15     3.0000    22.3333    2.9384  -19.3333      -2.33R
R denotes an obs. with a large st. resid.

Means for Ec7.0
Group7.0             Mean       Stdev
       1             22.25      1.944
       2             15.32      1.944
Time7.0
      1              19.12      1.799
      2              18.44      2.078
Group7.0*Time7.0
       1        1    22.17      2.545
       1        2    22.33      2.938
       2        1    16.08      2.545
       2        2    14.56      2.938
```

Appendix AK. Example Spreadsheet for Calculating the Test Statistic for the Friedman-Type Rank Test for H10

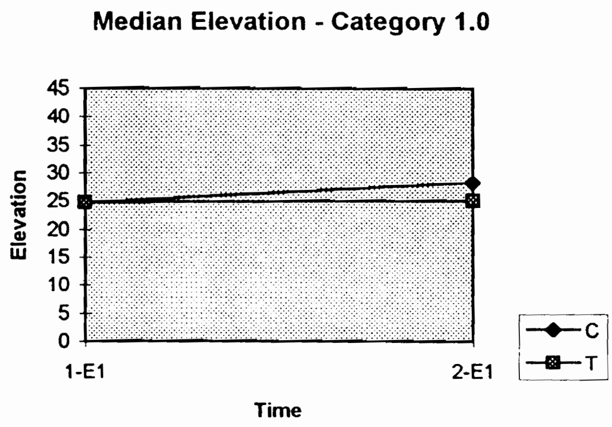
2 Factor Design for H10 Rank-Based Tests									
Category 2.0 - Elevation									
Calculations for COLUMN (time) effects									
			j = 1	TIME	**B (columns)**		*j = 2*		
		1st Evaluation			2nd Evaluation				
		n_{ijk}	Ranks			n_{ijk}	Ranks		
		1.67	1.5	n_{11} =		5.00	3.5	n_{12}=	$n_{1.}$=
		1.67	1.5	12		8.33	5	7	19
		5.00	3.5	mean =		10.00	6	mean =	
		15.00	8	22.78		11.67	7	18.33	
		18.33	9	sumranks=		25.00	10.5	sumranks=	
i=1	Control	25.00	10.5	127.0		28.33	12	63.0	
		31.67	13			40.00	19		
		35.00	16						
		35.00	16						
		35.00	16						
		35.00	16						
GROUP		35.00	16						
A (rows)		n_{ijk}	Ranks			n_{ijk}	Ranks		
		5.00	4.5	n_{21}=		1.67	1.5	n_{22}=	$n_{2.}$=
		8.33	7.5	11		1.67	1.5	9	20
		11.67	10	mean =		5.00	4.5	mean =	
		11.67	10	19.24		5.00	4.5	9.44	
i=2	Treatment	15.00	13	sumranks=		5.00	4.5	sumranks=	
		15.00	13	144.0		8.33	7.5	66.0	
		25.00	16			11.67	10		
		25.00	16			15.00	13		
		25.00	16			31.67	19		
		28.33	18						
		41.67	20						
modsumranks		R_1 =	13.884			R_2 =	6.616		
		$n_{.1}$=	23			$n_{.2}$=	16		
		$E_0(R_1)$ =	12.091			$E_0(R_2)$ =	8.409		
	R = (	1.793	-1.793	)'	Sigma =	0.82094	-0.82094		
						-0.82094	0.82094		
T = **R'Σ^-R**									
	R'			Σ^-			**R**		
T =	1.793	-1.793	*	0.304544	-0.30454	*	1.793		
				-0.304544	0.304544		-1.793		
T =	**3.918091**								

Group X Time Graphs of Mean Accuracy and Median Accuracy Continued

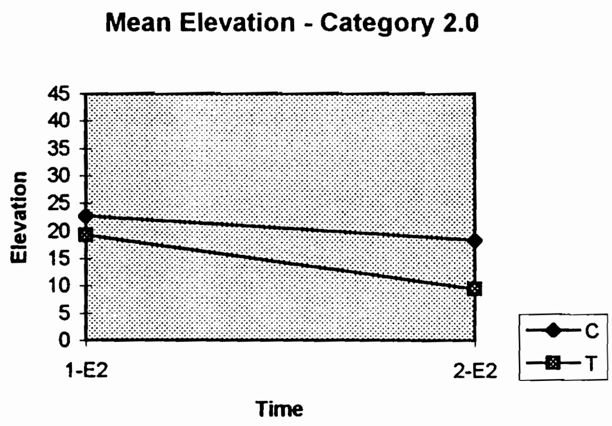
	Means	
Group	1-E1	2-E1
C	22.49	24.71
T	25.37	24.39



	Medians	
Group	1-E1	2-E1
C	24.7	28.3
T	24.7	25

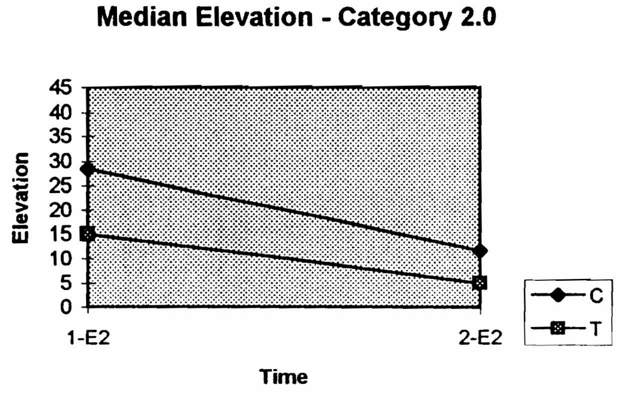


	Means	
Group	1-E2	2-E2
C	22.78	18.33
T	19.25	9.46

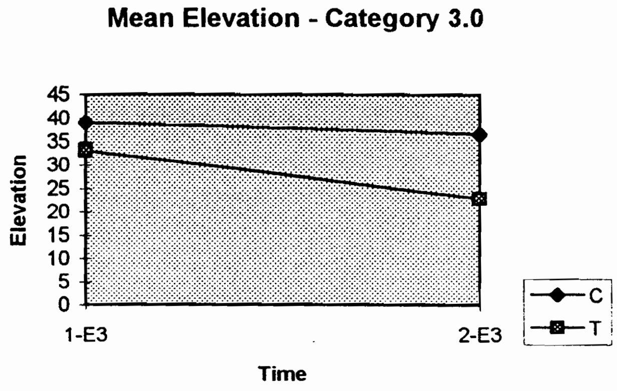


Group X Time Graphs of Mean Accuracy and Median Accuracy Continued

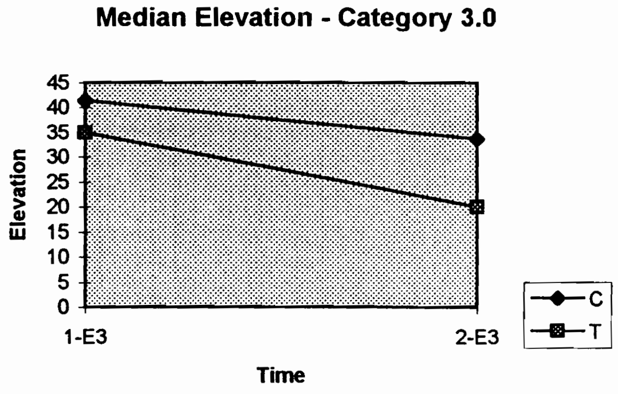
	Medians	
Group	1-E2	2-E2
C	28.35	11.7
T	15	5



	Means	
Group	1-E3	2-E3
C	39	36.65
T	33.08	22.78

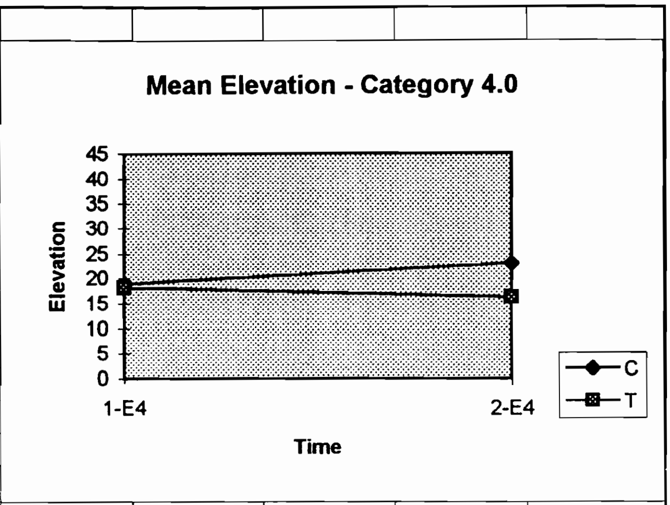


	Medians	
Group	1-E3	2-E3
C	41.5	33.75
T	35	20

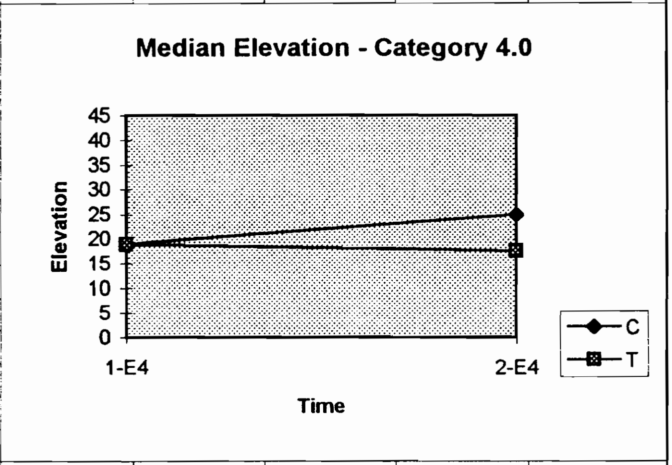


Group X Time Graphs of Mean Accuracy and Median Accuracy Continued

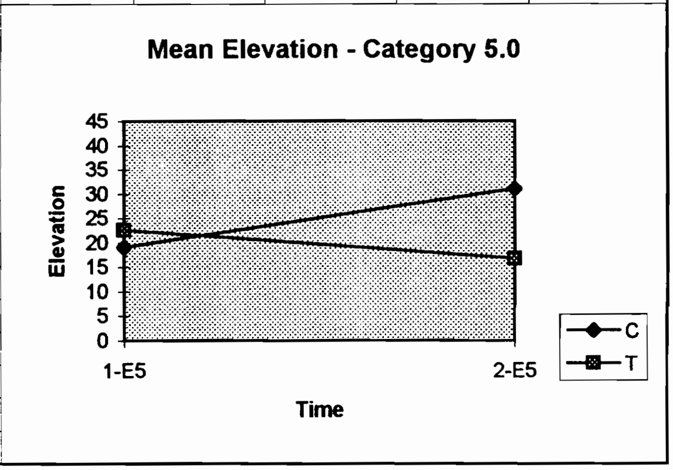
	Means	
Group	1-E4	2-E4
C	18.8	23.1
T	18.09	16.3



	Medians	
Group	1-E4	2-E4
C	18.8	25
T	18.8	17.55

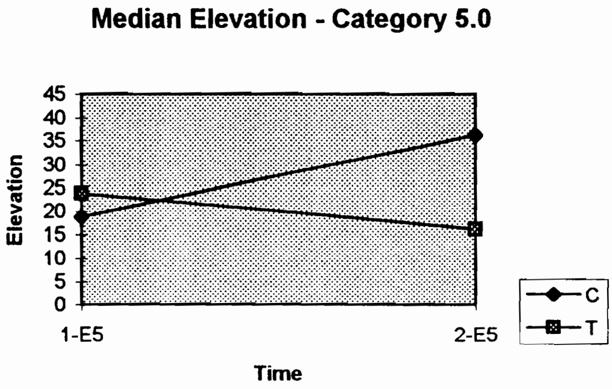


	Means	
Group	1-E5	2-E5
C	19.08	31.04
T	22.55	16.72

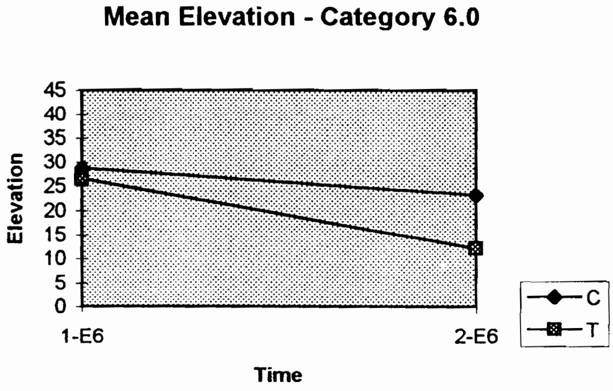


Group X Time Graphs of Mean Accuracy and Median Accuracy Continued

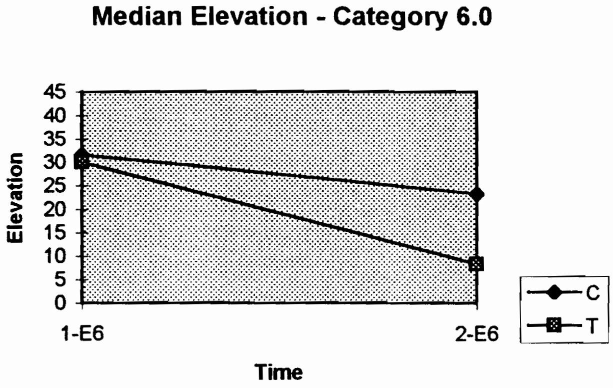
	Medians	
Group	1-E5	2-E5
C	18.8	36.3
T	23.8	16.3



	Means	
Group	1-E6	2-E6
C	28.67	23.19
T	26.64	12.03

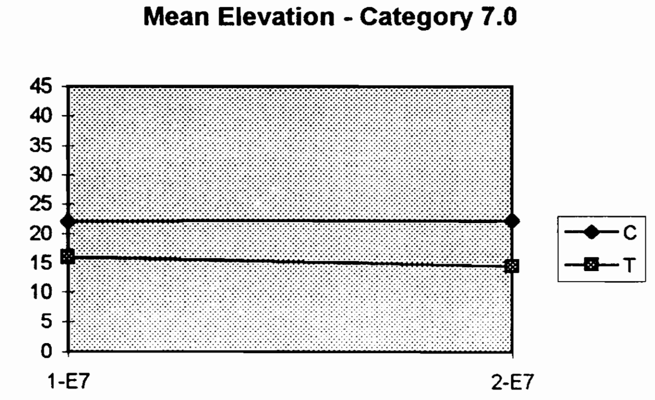


	Medians	
Group	1-E6	2-E6
C	31.7	23.3
T	30	8.3

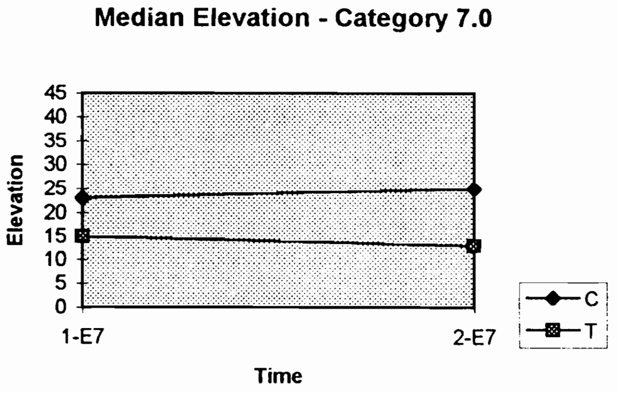


Group X Time Graphs of Mean Accuracy and Median Accuracy Continued

	Means	
Group	1-E7	2-E7
C	22.17	22.33
T	16.08	14.56

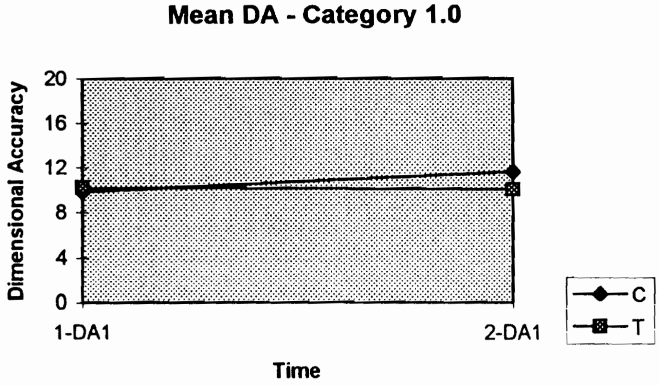


	Medians	
Group	1-E7	2-E7
C	23	25
T	15	13

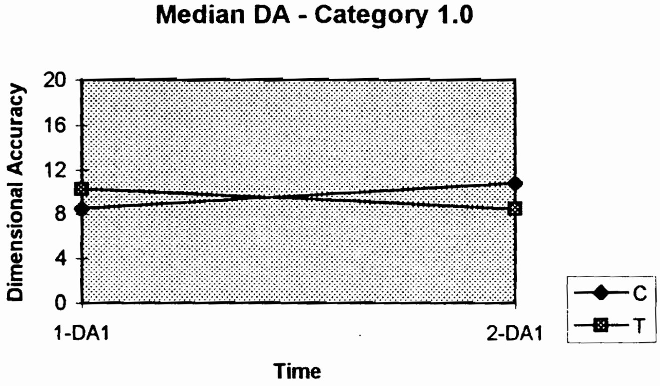


Group x Time Graphs of Mean Accuracy and Median Accuracy Continued

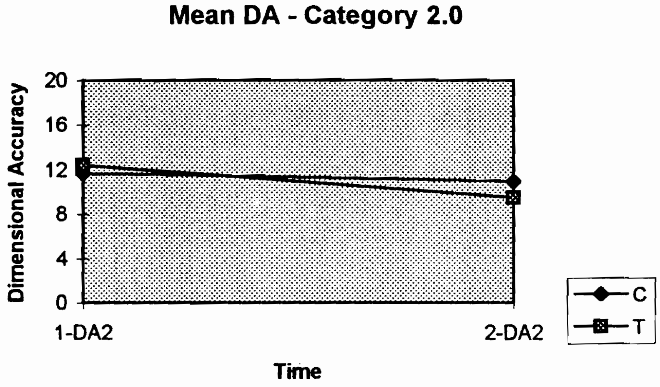
	Means	
Group	1-DA1	2-DA1
C	9.85	11.65
T	10.26	10.05



	Medians	
Group	1-DA1	2-DA1
C	8.5	10.8
T	10.3	8.5

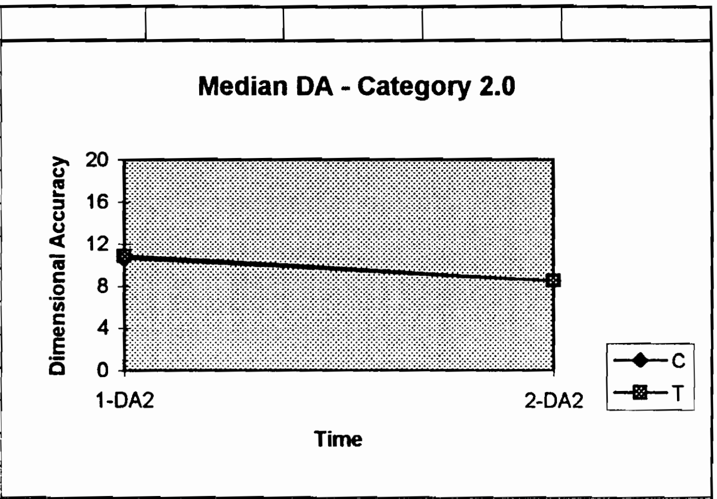


	Means	
Group	1-DA2	2-DA2
C	11.69	10.97
T	12.42	9.5

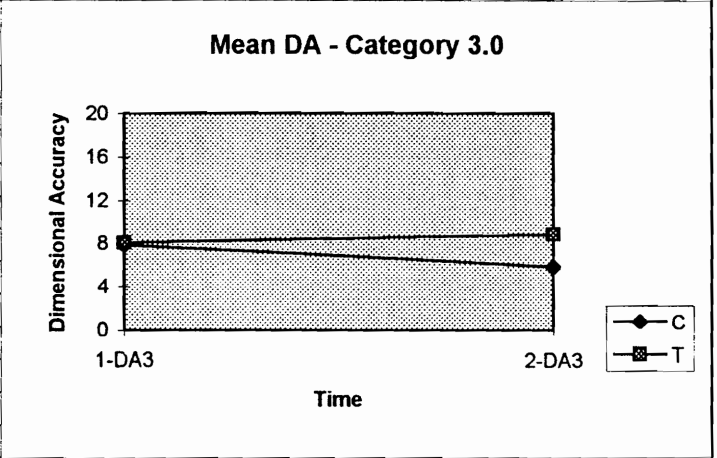


Group x Time Graphs of Mean Accuracy and Median Accuracy Continued

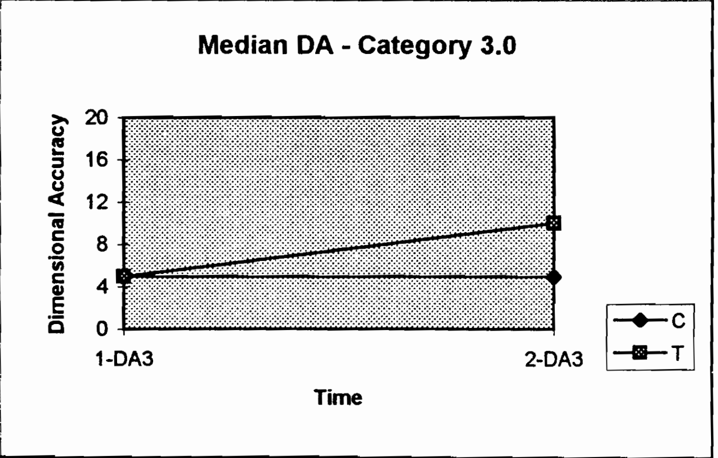
	Medians	
Group	1-DA2	2-DA2
C	10.55	8.5
T	10.8	8.5



	Means	
Group	1-DA3	2-DA3
C	7.82	5.85
T	8.08	8.89

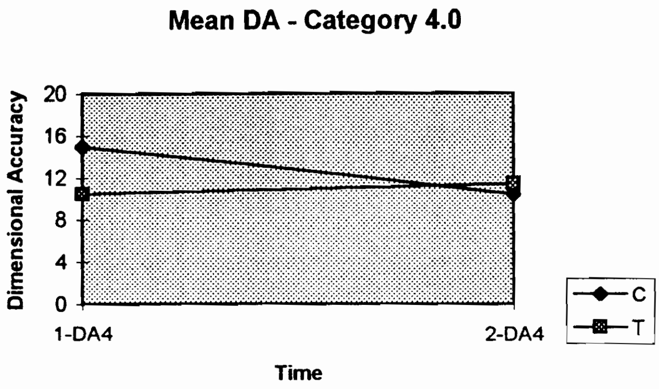


	Medians	
Group	1-DA3	2-DA3
C	5	5
T	5	10

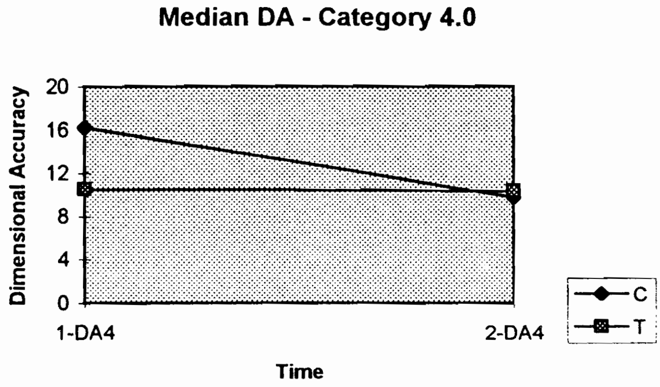


Group x Time Graphs of Mean Accuracy and Median Accuracy Continued

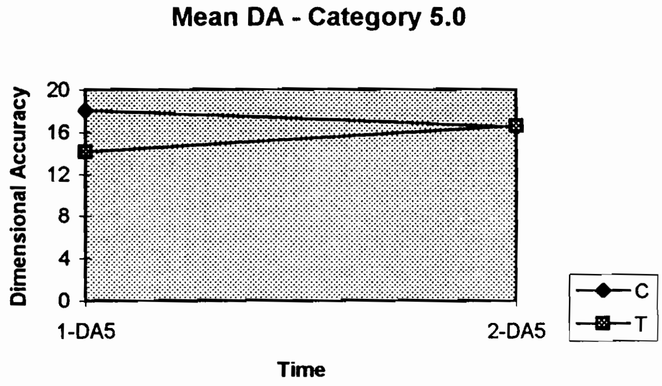
	Means	
Group	1-DA4	2-DA4
C	15	10.49
T	10.5	11.42



	Medians	
Group	1-DA4	2-DA4
C	16.2	9.8
T	10.55	10.4

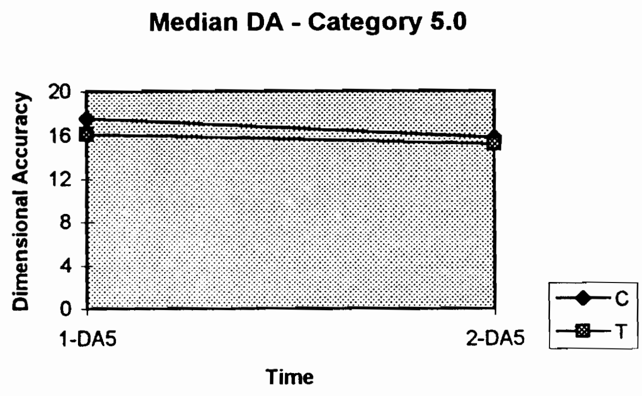


	Means	
Group	1-DA5	2-DA5
C	18.11	16.47
T	14.13	16.58

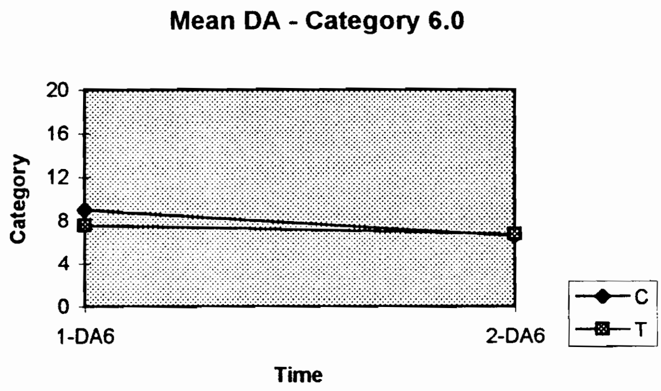


Group x Time Graphs of Mean Accuracy and Median Accuracy Continued

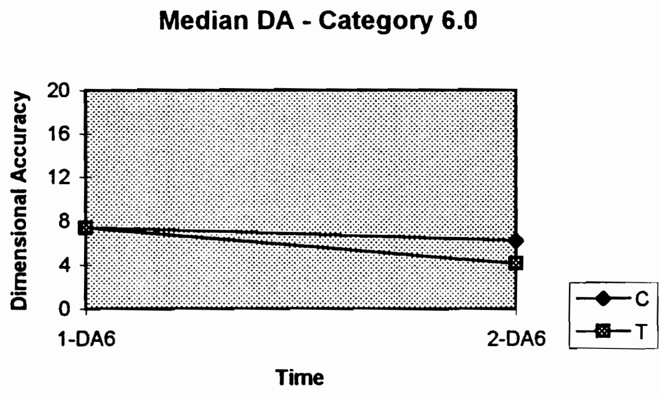
	Medians	
Group	1-DA5	2-DA5
C	17.5	15.75
T	16.1	15.15


Median DA - Category 5.0

	Means	
Group	1-DA6	2-DA6
C	8.99	6.54
T	7.53	6.73

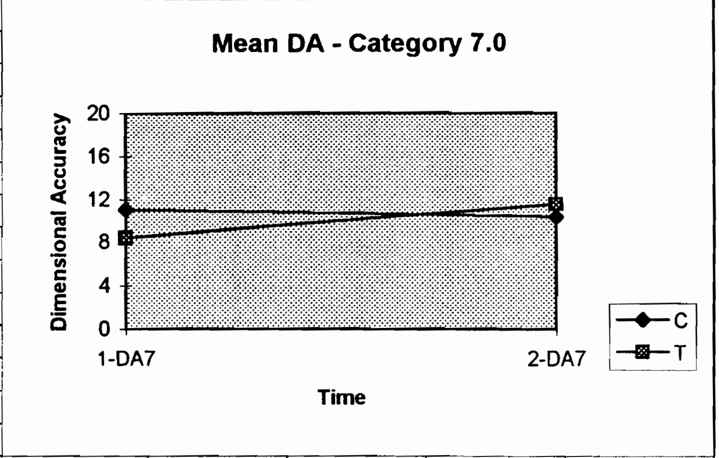

Mean DA - Category 6.0

	Medians	
Group	1-DA6	2-DA6
C	7.4	6.2
T	7.35	4.1

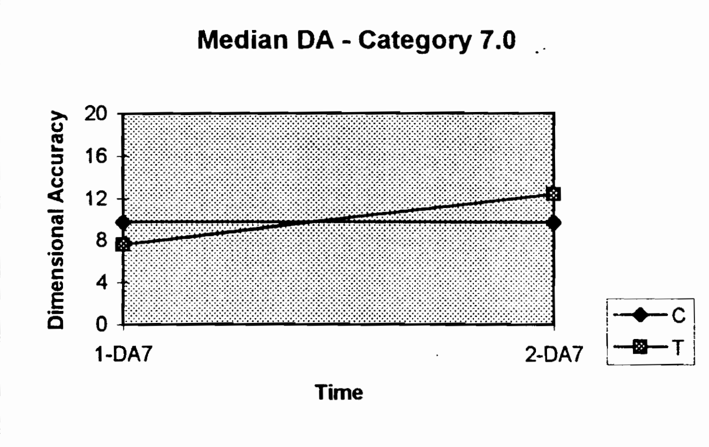

Median DA - Category 6.0

Group x Time Graphs of Mean Accuracy and Median Accuracy Continued

	Means	
Group	1-DA7	2-DA7
C	11.03	10.28
T	8.4	11.51



	Medians	
Group	1-DA7	2-DA7
C	9.7	9.7
T	7.6	12.4

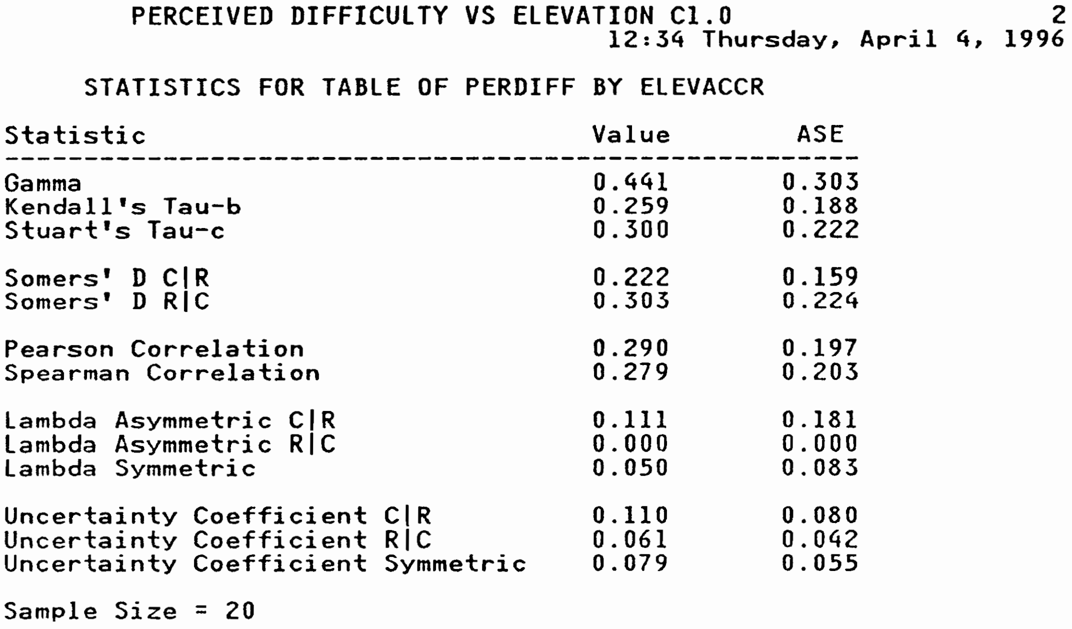


Appendix AM. Example SAS Output from H11 Ordered Categories Tests

Key: WST = worse accuracy; BST = better accuracy
E = easy; SE = somewhat easy; SD = somewhat difficult; D = difficult

```
PERCEIVED DIFFICULTY VS ELEVATION C1.0                          1
                                  12:34 Thursday, April 4, 1996

     TABLE OF PERDIFF BY ELEVACCR

 PERDIFF      ELEVACCR

 Frequency|
 Percent  |
 Row Pct  |
 Col Pct  |WST     |BST     |  Total
 ---------+--------+--------+
 E        |      2 |      1 |      3
          |  10.00 |   5.00 |  15.00
          |  66.67 |  33.33 |
          |  22.22 |   9.09 |
 ---------+--------+--------+
 SE       |      3 |      3 |      6
          |  15.00 |  15.00 |  30.00
          |  50.00 |  50.00 |
          |  33.33 |  27.27 |
 ---------+--------+--------+
 SD       |      4 |      5 |      9
          |  20.00 |  25.00 |  45.00
          |  44.44 |  55.56 |
          |  44.44 |  45.45 |
 ---------+--------+--------+
 D        |      0 |      2 |      2
          |   0.00 |  10.00 |  10.00
          |   0.00 | 100.00 |
          |   0.00 |  18.18 |
 ---------+--------+--------+
 Total           9       11       20
             45.00    55.00   100.00
```

Example SAS Output from H11 Ordered Categories Tests Continued



```
        PERCEIVED DIFFICULTY VS ELEVATION C1.0                   2
                                  12:34 Thursday, April 4, 1996

     STATISTICS FOR TABLE OF PERDIFF BY ELEVACCR

Statistic                                    Value        ASE
------------------------------------------------------------
Gamma                                        0.441       0.303
Kendall's Tau-b                              0.259       0.188
Stuart's Tau-c                               0.300       0.222

Somers' D C|R                                0.222       0.159
Somers' D R|C                                0.303       0.224

Pearson Correlation                          0.290       0.197
Spearman Correlation                         0.279       0.203

Lambda Asymmetric C|R                        0.111       0.181
Lambda Asymmetric R|C                        0.000       0.000
Lambda Symmetric                             0.050       0.083

Uncertainty Coefficient C|R                  0.110       0.080
Uncertainty Coefficient R|C                  0.061       0.042
Uncertainty Coefficient Symmetric            0.079       0.055

Sample Size = 20
```

Appendix AN. Test Statistics and Rejection Regions for H11 and H12

Test Statistic for H11

SAS was used to calculate the statistics necessary to compute the following test statistic.

$$Z_{obs} = \frac{\hat{\gamma}}{ASE(\hat{\gamma})}$$

Where gamma hat $(\hat{\gamma})$ is an estimate of gamma, the relationship between the two ordered category variables in the population and $ASE(\hat{\gamma})$ is the asymptotic standard error of gamma hat. These statistics were calculated by SAS for each test and used to compute Z_{obs}. Gamma hat is based on the difference between concordant and discordant pairs of observations within the ordered categories matrix as a proportion of the total number of concordant and discordant pairs (Schulman, 1994). Thus, a positive value of gamma hat results from more concordant pairs than discordant pairs and implies a positive relationship between the two ordered category variables. Appendix AM contains an example SAS output from the H11 ordered categories tests.

Rejection Region for H11

The test statistic, Z_{obs}, is the standardized version of gamma hat. Z_{obs} has an approximately normal distribution with a mean equal to zero and standard deviation of one. For the hypothesized negative relationship between perceived difficulty and accuracy, the null hypothesis was rejected whenever Z_{obs} was less than $Z_{0.05}$= -1.645 (Ott, 1984, p. 696). When the null hypothesis was rejected, there was a five percent or less chance that Z_{obs} would be this small (i.e., a large negative value) if the null hypothesis was true.

Test Statistic for H12

SAS was used to calculate the statistics necessary to compute the following test statistic.

$$Z_{obs} = \frac{\hat{\gamma}}{ASE(\hat{\gamma})}$$

Where gamma hat $(\hat{\gamma})$ is an estimate of gamma, the relationship between the two ordered category variables in the population and ASE $(\hat{\gamma})$ is the asymptotic standard error of gamma hat. These statistics were calculated by SAS for each test and used to compute Z_{obs}. Gamma hat is based on the difference between concordant and discordant pairs of observations within the ordered categories matrix as a proportion of the total number of concordant and discordant pairs (Schulman, 1994). Thus, a positive value of gamma hat results from more concordant pairs than discordant pairs and implies a positive relationship between the two ordered category variables.

Rejection Region for H12

The test statistic, Z_{obs}, is the standardized version of gamma hat. Z_{obs} has an approximately normal distribution with a mean equal to zero and standard deviation of one. For the hypothesized positive relationship between perceived accuracy and actual accuracy, the null hypothesis was rejected whenever Z_{obs} was greater than $Z_{0.05}$= 1.645 (Ott, 1984, p. 696). When the null hypothesis was rejected, there was a five percent or less chance that Z_{obs} would be this large if the null hypothesis was true.

Appendix AO. Descriptive Statistics of Variables used in Q12 Regression Analyses

```
MTB > Describe  'Eavg' 'DAavg'.
Descriptive Statistics (dependent variables)

Variable        N      Mean    Median    TrMean     StDev    SEMean
Eavg           67     22.83     21.70     22.73     10.99      1.34
DAavg          67    10.899     9.700    10.659     4.634     0.566

Variable      Min       Max        Q1        Q3
Eavg         2.40     45.90     15.40     31.00
DAavg       3.500    24.200     7.800    13.600


MTB > Describe 'qptrng'-'age'.
Descriptive Statistics (continuous independent variables)

Variable        N        N*      Mean    Median    TrMean     StDev    SEMean
qptrng         67         0      48.3       8.0      21.3     187.6      22.9
exp            67         0     7.837     5.500     7.361     7.089     0.866
yrsqcqa        67         0     1.083     0.000     0.709     2.475     0.302
supv           67         0      6.40      0.00      1.70     28.46      3.48
age            65         2    30.323    29.000    29.881     6.844     0.849

Variable      Min       Max        Q1        Q3
qptrng        0.0    1514.0       0.0      30.0
exp         0.000    26.000     2.500    12.000
yrsqcqa     0.000    15.000     0.000     1.000
supv         0.00    220.00      0.00      2.00
age        22.000    49.000    25.000    34.000


MTB > Describe 'degree' 'size'.
Descriptive Statistics (ordinal independent variables)

Variable        N        N*      Mean    Median    TrMean     StDev    SEMean
degree         67         0    2.1791    2.0000    2.1639    0.4237    0.0518
size           66         1     2.076     2.000     2.133     0.966     0.119

Variable      Min       Max        Q1        Q3
degree     1.0000    3.0000    2.0000    2.0000
size        0.000     3.000     1.750     3.000

MTB > Describe 'engreduc' 'assess' 'gndr'.
Descriptive Statistics (indicator or nominal independent variables)

Variable        N        N*      Mean    Median    TrMean     StDev    SEMean
engreduc       67         0    0.7910    1.0000    0.8197    0.4096    0.0500
assess         67         0    0.2090    0.0000    0.1803    0.4096    0.0500
gndr           66         1    0.2273    0.0000    0.2000    0.4223    0.0520

Variable      Min       Max        Q1        Q3
engreduc   0.0000    1.0000    1.0000    1.0000
assess     0.0000    1.0000    0.0000    0.0000
gndr       0.0000    1.0000    0.0000    0.0000
```

Descriptive Statistics of Variables used in Q12 Regression Analyses Continued

```
MTB > Descr.be 'exec'-'std'.
Descriptive Statistics (indicator independent variables describing
current job function)

Variable          N      Mean    Median    TrMean     StDev    SEMean
exec             67    0.1194    0.0000    0.0820    0.3267    0.0399
oper             67    0.1194    0.0000    0.0820    0.3267    0.0399
tech             67    0.4478    0.0000    0.4426    0.5010    0.0612
std              67    0.2836    0.0000    0.2623    0.4541    0.0555

Variable        Min       Max        Q1        Q3
exec         0.0000    1.0000    0.0000    0.0000
oper         0.0000    1.0000    0.0000    0.0000
tech         0.0000    1.0000    0.0000    1.0000
std          0.0000    1.0000    0.0000    1.0000


MTB > Describe 'mfg'-'stloc'.
Descriptive Statistics (indicator independent variables describing
subject's employer)

Variable          N      Mean    Median    TrMean     StDev    SEMean
mfg              67    0.1493    0.0000    0.1148    0.3590    0.0439
svc              67    0.2239    0.0000    0.1967    0.4200    0.0513
fed              67    0.3134    0.0000    0.2951    0.4674    0.0571
stloc            67    0.2239    0.0000    0.1967    0.4200    0.0513

Variable        Min       Max        Q1        Q3
mfg          0.0000    1.0000    0.0000    0.0000
svc          0.0000    1.0000    0.0000    0.0000
fed          0.0000    1.0000    0.0000    1.0000
stloc        0.0000    1.0000    0.0000    0.0000
```

Appendix AP. Edited Minitab Session Files for Proposed Regression Equations

Recommended regression equation with elevation as the dependent variable

```
MTB > # The following regression includes the three best predictors
MTB > # from the stepwise analyses.  degree2 deleted the single
MTB > # data point with less than a bachelors degree.  expE2 is the same
MTB > # as expE1 with the suppression of the observations omitted from
MTB > # qptrng by qptrng5E.  That is, those without any q/p training and
MTB > # the extreme outlier with 1514 days of q/p training.  This data
MTB > # set is believed to be representative of evaluators in training.

Note:  degree2 was later transformed to degreeE by subtracting two from
each data point.  This increased the constant to 21.683 without changing
the coefficients.

MTB > Name c65 = 'FITS1' c66 = 'RESI1'
MTB > Regress 'Eavg' 3 'engrdegE' 'degree2' 'expE2';
SUBC>   Fits 'FITS1';
SUBC>   Constant;
SUBC>   Residuals 'RESI1'.

Regression Analysis

The regression equation is
Eavg = 4.3 - 6.46 engrdegE + 8.68 degree2 + 0.527 expE2

42 cases used 25 cases contain missing values

Predictor        Coef        Stdev    t-ratio        p
Constant         4.33        10.46       0.41    0.681
engrdegE       -6.459        4.002      -1.61    0.115
degree2         8.676        3.917       2.22    0.033
expE2          0.5268       0.2779       1.90    0.066

s = 9.786        R-sq = 27.2%     R-sq(adj) = 21.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    3     1356.91      452.30      4.72    0.007
Error        38     3638.77       95.76
Total        41     4995.69

SOURCE       DF      SEQ SS
engrdegE      1      668.26
degree2       1      344.53
expE2         1      344.12

Unusual Observations
Obs. engrdegE       Eavg        Fit  Stdev.Fit    Residual     St.Resid
 20       1.00     44.30      24.71       3.48       19.59        2.14R

R denotes an obs. with a large st. resid.
```

Regression equations with dimensional accuracy as the dependent variable

```
MTB > # The following run produced the highest Rsq for DA (with a reasonable
sample size).

MTB > Regress 'DAavg' 7 'exp3D' 'std' 'yrsqcqa'  'stloc' 'gndr' 'size3D'  &
MTB >     'fed';
SUBC>   Constant.
Regression Analysis

The regression equation is
DAavg = 19.1 - 0.426 exp3D - 5.79 std + 0.506 yrsqcqa + 4.22 stloc - 2.40 gndr
           - 2.74 size3D + 3.27 fed

54 cases used 13 cases contain missing values

Predictor       Coef       Stdev    t-ratio        p
Constant      19.127       2.004       9.55    0.000
exp3D       -0.42644     0.09760      -4.37    0.000
std           -5.787       1.456      -3.98    0.000
yrsqcqa       0.5060      0.1925       2.63    0.012
stloc          4.216       1.489       2.83    0.007
gndr          -2.399       1.282      -1.87    0.068
size3D       -2.7436      0.9421      -2.91    0.006
fed            3.269       1.580       2.07    0.044

s = 3.366       R-sq = 47.9%     R-sq(adj) = 39.9%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     7      478.73       68.39      6.04    0.000
Error         46      521.11       11.33
Total         53      999.83

SOURCE        DF      SEQ SS
exp3D          1      164.82
std            1       63.12
yrsqcqa        1       63.32
stloc          1       43.86
gndr           1       47.53
size3D         1       47.57
fed            1       48.51

Unusual Observations
Obs.    exp3D      DAavg        Fit  Stdev.Fit    Residual    St.Resid
  5       2.5     21.900     15.317      1.250       6.583       2.11R
 41      18.0     13.600     14.079      2.638      -0.479      -0.23 X
 62      15.0     12.900      1.981      1.748      10.919       3.80R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

<u>The next run drops stloc and produces what may be the most practical equation when DA is the dependent variable.</u>

```
MTB > Regress 'DAavg' 6 'exp3D' 'std' 'yrsqcqa' 'size3D' 'mfg' 'svc';
SUBC>   Constant.
Regression Analysis

The regression equation is
DAavg = 22.8 - 0.421 exp3D - 4.64 std + 0.532 yrsqcqa - 3.07 size3D - 2.58 mfg
           - 4.76 svc

54 cases used 13 cases contain missing values

Predictor        Coef        Stdev    t-ratio        p
Constant       22.818        2.533       9.01    0.000
exp3D        -0.42075      0.09573      -4.40    0.000
std            -4.639        1.274      -3.64    0.001
yrsqcqa        0.5317       0.1938       2.74    0.009
size3D        -3.0718       0.8707      -3.53    0.001
mfg            -2.576        1.593      -1.62    0.113
svc            -4.759        1.431      -3.33    0.002

s = 3.389       R-sq = 46.0%     R-sq(adj) = 39.1%

Analysis of Variance

SOURCE       DF         SS          MS         F        p
Regression    6     460.09       76.68      6.68    0.000
Error        47     539.74       11.48
Total        53     999.83

SOURCE       DF     SEQ SS
exp3D         1     164.82
std           1      63.12
yrsqcqa       1      63.32
size3D        1      41.76
mfg           1       0.09
svc           1     126.99

Unusual Observations
Obs.    exp3D       DAavg        Fit  Stdev.Fit    Residual     St.Resid
  5       2.5      21.900     13.935      1.230       7.965        2.52R
 41      18.0      13.600     14.005      2.651      -0.405       -0.19 X
 62      15.0      12.900      2.652      1.554      10.248        3.40R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

MTB > # Note that when stloc was dropped from stepwise, fed was replaced by
MTB > # mfg and service, and gndr disappeared from the equation.
```

Recommended regression equation with dimensional accuracy as the dependent variable

NOTE: The following run drops the obs. 62 (extremely large standard residual in the previous runs). The variables are relabeled, but contain the same data as described earlier.

```
MTB > Regress 'DAavg' 6 'exp9D'-'size9D';
SUBC>   Constant.

Regression Analysis

The regression equation is
DAavg = 25.3 - 0.525 exp9D - 6.29 std9 - 2.80 mfg9 - 5.23 svc9
       + 0.631 yrsqcqa9 - 3.73 size9D

53 cases used 14 cases contain missing values

Predictor        Coef        Stdev     t-ratio        p
Constant       25.289        2.312       10.94    0.000
exp9D        -0.52509      0.08820       -5.95    0.000
std9           -6.292        1.196       -5.26    0.000
mfg9           -2.805        1.399       -2.00    0.051
svc9           -5.230        1.262       -4.14    0.000
yrsqcqa9       0.6309       0.1720        3.67    0.001
size9D        -3.7324       0.7828       -4.77    0.000

s = 2.974       R-sq = 59.1%     R-sq(adj) = 53.8%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     6     588.373      98.062     11.09    0.000
Error         46     406.752       8.842
Total         52     995.125

SOURCE        DF      SEQ SS
exp9D          1     183.174
std9           1      99.970
mfg9           1       2.715
svc9           1      17.530
yrsqcqa9       1      83.946
size9D         1     201.038

Unusual Observations
Obs.    exp9D       DAavg         Fit  Stdev.Fit    Residual     St.Resid
  5       2.5      21.900      15.014      1.115       6.886        2.50R
 27       2.0      16.300      10.481      0.830       5.819        2.04R
 41      18.0      13.600      14.104      2.327      -0.504       -0.27 X

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

Appendix AQ. Examples from the Table used to Summarize Qualitative Comments (First Evaluation)

ID#	Item	(+/++)	Area to Address	(+) Strengths	(-/--)	Area to Address	(-) Areas for Improvement	Site Visit Issues
1446	1.1		a(1)	CFI clearly states its values. Senior executives spend a good deal of time with customers and direct reports to reinforce these values.		a(1)	Focus is principally upon the customer. Other stakeholders (community, public, suppliers, stockholders) receive little or no mention.	None
			a(2)	CFI's strategic plan is grounded in five key strategies designed to create objectives that continuously challenge leadership to excel.		a(2)	Does not appear to address the possibility of marketplace erosion in military-related businesses.	
			a(3)	Company performance is reviewed through the use of feedback from employees, internal assessments, customer feedback, and third-party surveys. Specialists are called in when performance lags.		a(3)	Performance is spoken of in very general terms. Are financial, hr, quality performance all handled as described?	
			b	A 360° review process is used to evaluate senior managers annually. Assessment instruments are used to evaluate managerial effectiveness.		b	Annual reviews of senior executives may not be often enough.	
1513	1.1	++		Quarterly data gathering meetings with customers, suppliers, and partners	-		Involve upper management directly in worker's careers (all-hands mtgs, skip levels, etc.)	Survey workers for perception of management involvement
		+		Definition of mission				
		+		Definition of values	-		Start implementation planning from the start of corporate planning	
		++		Monitor feedback from employees, reviews, internal assessment, surveys, and training				
		++	a(2)	Formalization of business model				

Appendix AR. Experts' Qualitative Comments for Item 1.1

1.1 SENIOR EXECUTIVE LEADERSHIP (45 points)

+/++	*Area to Address*	*(+) Strengths*
+	a	A company credo, originated by the founders, forms the basis for CFI's mission and values, and it is communicated to every part of the organization.
+	a	The senior executive team sets the standards of performance and behavior for the rest of the company.
+	a	Senior executives meet routinely (quarterly visits) with customers, suppliers, and partners around the world.
++	a	The core values of the company are amplified and clarified for all employees by translation of them to explicit statements of values and the required behavior.
+	a	Key business strategies are consistent with and related to the core values.
+	a	All managers are regularly retrained in guidelines for measuring employee performance relative to the core values.
+	a	Senior executives are active leaders of processes and studies to improve both customer relationships and understanding of competitive and market environments.
+	a	Senior executives are intimately involved in the measurement, assessment, and review of all facets of CFI operations. This includes providing assistance and additional company resources to units not performing as planned.
+	b	Senior executives use 360 degree feedback, employee surveys, and customer feedback to monitor leadership effectiveness, and they have made refinements such as simplification of the core value descriptions.

-/--	*Area to Address*	*(-) Areas for Improvement*
-	a	It is unclear what the specific roles, responsibilities, and actions of the senior executive team have been in reinforcing values, establishing business strategies, and reviewing performances.
-	a	There is no evidence that values and expectations take into account all stakeholders; the interests of the public and the community are not cited.
-	a	Senior executive leadership approaches are not described in sufficient detail to assess effectiveness.
-	b	It is unclear to what extent there have been cycles of improvement in the leadership system.

Site Visit Issues:

- Verify the extent of deployment of the mission credo, core values, and behaviors.
- Clarify the extent of senior executive visibility in leadership activities.
- Clarify to what extent cycles of improvement have taken place to improve senior executive effectiveness.
- Verify the use of the evaluation tools such as "System Effectiveness Assessment."
- Review the employee and external surveys used to measure adherence to the values and to customer requirements.

Appendix AS. Count Data from Q5 Content Analysis

Control Group 1st Evaluation, Category 1.0 - Leadership

	Item 1.1					Areas for			Site Visit			Item 1.1		Item 1.2					Areas for			Site Visit			Item 1.2		Item 1.3					Areas for			Site Visit			Item 1.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1409	6	0	1	86%	67%	2	0	2				73%	44%	3	0	1	75%	60%	0	1	0				60%	22%	4	0	1	80%	80%	0	0	2				57%	25%
1413	2	2	2	33%	33%							33%	17%	4	0	0	100%	80%	1	0	0				100%	31%	2	0	0	100%	40%	1	0	1				75%	19%
1425	1	2	1	25%	22%	1	1	0	0	1	0	29%	22%	3	0	1	75%	60%	0	1	2				43%	22%	2	1	0	67%	50%	0	1	2				33%	19%
1436	6	2	1	67%	78%	1	0	1	1	0	1	62%	50%	5	0	2	71%	100%	1	0	0	1	0	1	70%	44%	5	0	4	56%	100%	1	0	1	0	1	1	46%	41%
1446	3	1	0	75%	39%	1	1	2				50%	28%	4	1	0	80%	90%	0	1	0				67%	31%	3	0	1	75%	60%	1	1	1				57%	28%
1448	2	3	2	29%	39%				0	0	1	25%	19%	2	1	0	67%	50%	0	1	1	0	1	0	33%	22%	1	1	0	50%	30%	0	0	1	0	0	1	25%	9%
1503	4	1	1	67%	50%							67%	25%	1	0	0	100%	20%	0	1	0				50%	9%													
1506	4	1	0	80%	50%	1	2	2	2	1	1	50%	50%	2	0	0	100%	40%	0	1	1	1	0	0	60%	22%	1	0	1	50%	20%	2	0	0				75%	19%
1513	3	0	1	75%	33%	0	1	1	0	1	0	43%	22%	3	0	0	100%	60%	0	1	1	1	0	1	57%	28%	4	0	1	80%	80%	1	0	2	0	1	0	56%	34%
1528	3	3	1	43%	50%	1	1	2				36%	33%	3	0	0	100%	60%	0	0	1				75%	19%	1	0	0	100%	20%	0	1	0				50%	9%
1530	3	1	0	75%	39%	0	2	1				43%	25%	2	1	0	67%	50%	1	0	1				60%	22%	2	0	1	67%	40%	1	0	1				60%	19%
1555	4	0	0	100%	44%	1	0	0				100%	28%	1	1	1	33%	30%	2	0	0	0	1	0	50%	25%	4	0	1	80%	80%	1	0	0				83%	31%
Avgs.	**3.4**	**1.3**	**0.8**	**63%**	**45%**	**0.9**	**0.9**	**1.2**	**0.6**	**0.6**	**0.6**	**51%**	**30%**	**2.8**	**0.3**	**0.4**	**81%**	**58%**	**0.4**	**0.6**	**0.6**	**0.6**	**0.4**	**0.4**	**60%**	**25%**	**2.6**	**0.2**	**0.9**	**73%**	**55%**	**0.7**	**0.3**	**1**	**0**	**0.7**	**0.7**	**56%**	**23%**
Experts	**9**					**4**			**5**					**5**					**6**			**5**					**5**					**6**			**5**				

Treatment Group 1st Evaluation, Category 1.0 - Leadership

	Item 1.1					Areas for			Site Visit			Item 1.1		Item 1.2					Areas for			Site Visit			Item 1.2		Item 1.3					Areas for			Site Visit			Item 1.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2401	6	2	3	55%	78%	1	0	1	1	1	1	50%	53%	4	0	0	100%	80%	0	0	0	0	1	0	80%	28%	5	0	0	100%	100%	2	1	2	0	1	2	54%	50%
2411	3	2	0	60%	44%	0	0	1	0	0	1	43%	22%	2	1	0	67%	50%	0	1	0				50%	19%	3	0	1	75%	60%							75%	19%
2424	3	2	1	50%	44%	0	1	2				33%	25%	1	2	0	33%	40%	0	1	1	0	1	0	17%	19%	1	1	2	25%	30%	1	0	1				33%	16%
2449	4	1	1	67%	50%	1	0	0	1	0	0	75%	36%	2	0	1	67%	40%	1	1	0	2	0	0	71%	34%	3	0	1	75%	60%	0	1	0	1	0	0	67%	28%
2505	3	1	1	60%	39%	0	1	1				43%	22%	1	2	0	33%	40%	1	1	1				33%	22%	2	0	0	100%	40%	1	0	1				75%	19%
2508	4	2	2	50%	56%	1	0	4				38%	33%	4	1	0	80%	90%	2	1	1				67%	44%	5	0	3	63%	100%	0	1	1				50%	34%
2515	3	2	2	43%	44%	0	1	5	0	1	2	19%	28%	1	1	0	50%	30%	1	0	2	0	1	1	29%	19%	2	0	2	50%	40%	1	0	2	1	0	2	40%	25%
2531	4	3	2	44%	61%	0	1	2				33%	33%	3	1	0	75%	70%							75%	22%	3	0	1	75%	60%	0	0	2				50%	19%
2536	2	1	1	50%	28%	0	0	1				40%	14%	0	2	0	0%	20%	0	0	1				0%	6%	2	0	0	100%	40%				0	0	1	67%	13%
2537	5	2	0	71%	67%	1	0	0				75%	39%	3	0	0	100%	60%	2	0	0	2	0	0	100%	44%	4	0	1	80%	80%	1	0	1	0	0	1	63%	31%
2545	6	1	1	75%	72%	0	0	1	1	0	3	54%	42%	3	0	0	100%	60%	0	0	1	2	0	0	83%	31%	3	0	0	100%	60%	1	0	0				100%	25%
2546	5	2	3	50%	67%	0	0	3				38%	33%	4	0	1	80%	80%	2	0	1				75%	38%	4	0	2	67%	80%	1	0	1				63%	31%
2548	1	1	3	20%	17%	1	0	3				22%	14%	1	3	0	25%	50%	0	3	1				13%	25%	3	1	2	50%	70%	2	0	4				42%	34%
Avgs.	**3.8**	**1.7**	**1.5**	**53%**	**51%**	**0.4**	**0.3**	**1.8**	**0.6**	**0.4**	**1.4**	**43%**	**30%**	**2.2**	**1**	**0.2**	**62%**	**55%**	**0.8**	**0.7**	**0.8**	**1**	**0.5**	**0.2**	**53%**	**27%**	**3.1**	**0.2**	**1.2**	**74%**	**63%**	**0.9**	**0.3**	**1.4**	**0.4**	**0.2**	**1.2**	**60%**	**26%**
Experts	**9**					**4**			**5**					**5**					**6**			**5**					**5**					**6**			**5**				
Overall																																							
Avgs.	**3.6**	**1.5**	**1.2**	**58%**	**48%**	**0.6**	**0.5**	**1.6**	**0.6**	**0.5**	**1.0**	**47%**	**30%**	**2.5**	**0.7**	**0.3**	**71%**	**56%**	**0.6**	**0.6**	**0.7**	**0.8**	**0.5**	**0.3**	**57%**	**26%**	**2.9**	**0.2**	**1**	**73%**	**59%**	**0.8**	**0.3**	**1.2**	**0.3**	**0.4**	**1**	**58%**	**25%**

Count Data from Q5 Content Analysis for Category 2.0

Control Group 1st Evaluation, Category 2.0 - Information & Analysis																																							
	Item 2.1					Areas for			Site Visit			Item 2.1		Item 2.2					Areas for			Site Visit			Item 2.2		Item 2.3					Areas for			Site Visit			Item 2.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1409	3	1	6	30%	70%	1	0	3				29%	38%	2	1	1	50%	83%	0	0	3				29%	28%	2	1	1	50%	63%	0	1	0				40%	27%
1413	2	0	1	67%	40%							67%	17%	2	0	0	100%	67%							100%	22%	0	0	1	0%	0%							0%	0%
1417	3	1	2	50%	70%							50%	29%	1	1	6	13%	50%							13%	17%	2	1	1	50%	63%							50%	23%
1431	1	1	3	20%	30%	0	0	1				17%	13%	1	0	1	50%	33%	0	0	1				33%	11%	1	2	1	25%	50%							25%	18%
1436	4	0	9	31%	80%	1	1	0	0	0	2	29%	46%	2	1	7	20%	83%	0	0	1	1	0	1	23%	39%	2	1	6	22%	63%	1	1	0	1	0	0	33%	45%
1442	5	0	2	71%	100%	0	0	1				63%	42%	1	0	3	25%	33%	0	1	2				14%	17%	3	0	1	75%	75%							75%	27%
1446	2	2	2	33%	60%	1	0	1				38%	33%	2	0	2	50%	67%	1	0	0				60%	33%	2	1	1	50%	63%	0	0	4				25%	23%
1448	2	2	1	40%	60%	0	0	2	0	1	0	25%	29%	1	1	3	20%	50%	0	0	2				14%	17%	0	2	0	0%	25%				0	0	1	0%	9%
1506	2	1	2	40%	50%	0	0	2	0	0	2	22%	21%	0	2	1	0%	33%	0	0	2	0	0	2	0%	11%	0	0	1	0%	0%	0	1	1	0	0	1	0%	5%
1541	0	0	3	0%	0%	1	0	0				25%	8%	2	0	1	67%	67%							67%	22%	0	1	0	0%	13%	1	0	0				50%	14%
1552	1	0	1	50%	20%	0	0	2	0	0	2	17%	8%	0	1	1	0%	17%	0	0	2	0	0	1	0%	6%	0	1	1	0%	13%	1	0	1	0	0	1	20%	14%
1555	2	1	7	20%	50%				0	0	1	18%	21%	2	0	1	67%	67%	0	0	4				29%	22%	2	1	4	29%	63%	0	1	0				25%	27%
Avgs.	**2.3**	**0.8**	**3.3**	**38%**	**53%**	**0.4**	**0.1**	**1.3**	**0.0**	**0.2**	**1.4**	**33%**	**25%**	**1.3**	**0.6**	**2.3**	**38%**	**54%**	**0.1**	**0.1**	**1.9**	**0.3**	**0.0**	**1.3**	**32%**	**20%**	**1.2**	**0.9**	**1.5**	**25%**	**41%**	**0.4**	**0.6**	**0.9**	**0.3**	**0.0**	**0.8**	**29%**	**19%**
Experts	**5**					**4**			**3**					**3**					**3**			**3**					**4**					**3**			**4**				
Treatment Group 1st Evaluation, Category 2.0 - Information & Analysis																																							
	Item 2.1					Areas for			Site Visit			Item 2.1		Item 2.2					Areas for			Site Visit			Item 2.2		Item 2.3					Areas for			Site Visit			Item 2.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2405	2	0	4	33%	40%	0	0	3	0	0	1	20%	17%	1	1	3	20%	50%	0	0	3	0	1	1	10%	22%	0	2	1	0%	25%	0	0	3	0	0	1	0%	9%
2411	1	1	2	25%	30%				0	1	0	20%	17%	1	0	4	20%	33%							20%	11%	2	0	1	67%	50%				0	0	1	50%	18%
2414	1	3	6	10%	50%	2	1	0	1	0	2	25%	50%	2	1	4	29%	83%	2	0	2	1	1	1	36%	67%	2	0	2	50%	50%	0	1	1	0	1	1	25%	27%
2419	2	0	3	40%	40%	0	1	2	1	0	0	33%	29%	2	1	1	50%	83%	0	0	3	0	1	0	25%	33%	1	0	2	33%	25%	2	0	1	0	0	1	43%	27%
2424	0	1	3	0%	10%	0	1	1				0%	8%	0	1	2	0%	17%							0%	6%	0	2	4	0%	25%	1	0	1				13%	18%
2432	2	0	3	40%	40%	1	0	1	0	0	1	38%	25%	1	1	1	33%	50%	1	0	2				33%	28%	1	1	2	25%	38%	1	0	0				40%	23%
2433	1	2	2	20%	40%	1	0	2				25%	25%	1	1	3	20%	50%	1	0	1				29%	28%	2	0	2	50%	50%	2	0	0				67%	36%
2437	3	0	1	75%	60%	0	0	1				60%	25%	0	0	2	0%	0%	0	0	4				0%	0%	0	0	2	0%	0%	1	0	1				25%	9%
2449	2	1	3	33%	50%	0	0	4	0	1	0	18%	25%	0	0	3	0%	0%	1	0	2	0	1	1	13%	17%	3	0	1	75%	75%	0	0	1				60%	27%
2504	1	1	5	14%	30%	1	0	1				22%	21%	1	0	3	25%	33%	1	0	3				25%	22%	0	0	1	0%	0%	1	1	0				33%	14%
2507	2	1	4	29%	50%							29%	21%	1	0	1	50%	33%	1	0	2				40%	22%	0	2	2	0%	25%	1	0	0				20%	18%
2508	3	1	4	38%	70%	0	0	2				30%	29%	2	0	7	22%	67%	0	0	2				18%	22%	2	0	4	33%	50%	0	0	2				25%	18%
Avgs.	**1.7**	**0.9**	**3.3**	**30%**	**43%**	**0.5**	**0.3**	**1.7**	**0.3**	**0.3**	**0.7**	**27%**	**24%**	**1.0**	**0.5**	**2.8**	**22%**	**42%**	**0.7**	**0.0**	**2.4**	**0.3**	**1.0**	**0.8**	**21%**	**23%**	**1.1**	**0.6**	**2.0**	**28%**	**34%**	**0.8**	**0.2**	**0.9**	**0.0**	**0.3**	**1.0**	**33%**	**20%**
Experts	**5**					**4**			**3**					**3**					**3**			**3**					**4**					**3**			**4**				
Overall	**2.0**	**0.8**	**3.3**	**34%**	**48%**	**0.5**	**0.2**	**1.5**	**0.2**	**0.3**	**1.0**	**30%**	**25%**	**1.2**	**0.5**	**2.5**	**30%**	**48%**	**0.4**	**0.1**	**2.2**	**0.3**	**0.6**	**1.0**	**26%**	**22%**	**1.1**	**0.8**	**1.8**	**26%**	**38%**	**0.7**	**0.3**	**0.9**	**0.1**	**0.1**	**0.9**	**31%**	**20%**
Avgs.																																							

Count Data from Q5 Content Analysis for Category 3.0

Control Group 1st Evaluation, Category 3.0 - Strategic Planning																										
	Item 3.1					Areas for			Site Visit			Item 3.1		Item 3.2					Areas for			Site Visit			Item 3.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1410	2	0	2	50%	50%	1	0	1	0	0	1	43%	19%	1	1	1	33%	38%				0	0	1	25%	13%
1415	2	0	1	67%	50%	1	0	0				75%	19%	1	2	0	33%	50%	0	2	1				17%	25%
1425	2	1	1	50%	63%	0	1	2				29%	19%	1	2	3	17%	50%	2	0	1				33%	33%
1441	1	0	2	33%	25%	1	0	1				40%	13%	1	2	0	33%	50%	1	0	1				40%	25%
1443	1	0	2	33%	25%	2	0	2	1	0	0	50%	25%	1	1	1	33%	38%	2	1	0				50%	33%
1501	3	0	3	50%	75%							50%	19%	2	1	2	40%	63%							40%	21%
1522	1	1	2	25%	38%	0	1	0				20%	13%	0	1	0	0%	13%							0%	4%
1525	2	0	2	50%	50%	2	1	2	1	0	0	50%	34%	1	1	0	50%	38%	1	1	3	0	1	1	22%	29%
1527	1	0	2	33%	25%	1	0	2	0	0	2	25%	13%	0	1	1	0%	13%	0	0	2	0	0	1	0%	4%
1541	2	0	1	67%	50%							67%	13%	1	1	1	33%	38%							33%	13%
1554	0	1	1	0%	13%	0	1	2	0	1	1	0%	9%	0	0	1	0%	0%	0	1	0	0	1	0	0%	8%
Avgs.	1.5	0.3	1.7	42%	42%	0.9	0.4	1.3	0.4	0.2	0.8	41%	18%	0.8	1.2	0.9	25%	35%	0.9	0.7	1.1	0.0	0.5	0.8	24%	19%
Experts	4					6			6					4					4			4				
Treatment Group 1st Evaluation, Category 3.0 - Strategic Planning																										
	Item 3.1					Areas for			Site Visit			Item 3.1		Item 3.2					Areas for			Site Visit			Item 3.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2402	0	1	4	0%	13%	0	0	3				0%	3%	0	0	5	0%	0%	0	1	3				0%	4%
2407	1	1	4	17%	38%	1	0	0	0	0	1	25%	16%	1	0	0	100%	25%	3	0	0	2	0	1	86%	50%
2416	2	1	0	67%	63%	2	0	0				80%	28%	0	0	1	0%	0%	2	0	1				50%	17%
2423	2	0	2	50%	50%	0	2	3	0	0	2	18%	19%	1	0	1	50%	25%	1	2	2	1	0	2	30%	33%
2427	0	0	1	0%	0%	0	0	3	1	0	0	20%	6%	0	0	2	0%	0%	2	1	0	0	0	1	33%	21%
2429	0	1	2	0%	13%	0	0	1	0	0	1	0%	3%	1	0	3	25%	25%				1	0	0	40%	17%
2437	1	1	2	25%	38%	0	1	0	0	1	0	17%	16%	1	0	0	100%	25%							100%	8%
2439	2	1	1	50%	63%	0	0	1	0	0	2	29%	16%	1	0	3	25%	25%	0	0	1				20%	8%
2521	1	1	0	50%	38%	0	1	3	0	0	1	14%	13%	1	1	0	50%	38%	0	1	2	0	0	1	17%	17%
2532	0	0	3	0%	0%	0	0	3				0%	0%	0	1	2	0%	13%	1	1	0				20%	17%
2537	2	0	0	100%	50%	1	1	2	0	0	2	38%	22%	1	2	0	33%	50%	1	0	1	0	0	1	33%	25%
2551	0	1	2	0%	13%	0	0	4	0	0	2	0%	3%	0	1	0	0%	13%	1	0	2				25%	13%
2553	2	1	1	50%	63%	0	0	1				40%	16%	1	1	2	25%	38%	1	0	0				40%	21%
Avgs.	1.0	0.7	1.7	31%	34%	0.3	0.4	1.8	0.1	0.1	1.2	22%	12%	0.6	0.5	1.5	31%	21%	1.1	0.5	1.1	0.7	0.0	1.0	38%	19%
Experts	4					6			6					4					4			4				
Overall	1.3	0.5	1.7	36%	38%	0.5	0.4	1.6	0.2	0.1	1.1	30%	15%	0.7	0.8	1.2	28%	28%	1.0	0.6	1.1	0.4	0.2	0.9	31%	19%
Avgs.																										

Count Data from Q5 Content Analysis for Category 4.0

Control Group 1st Evaluation, Category 4.0 - Human Resource Development & Management																										
	Item 4.1					Areas for			Site Visit			Item 4.1		Item 4.2					Areas for			Site Visit			Item 4.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	0	0	9	0%	0%	0	1	1	0	0	2	0%	4%	2	0	3	40%	50%	0	0	1				33%	18%
1410	1	2	2	20%	67%	0	0	1	0	0	1	14%	15%	2	1	3	33%	63%	0	0	1	0	0	1	25%	23%
1415	0	0	3	0%	0%	1	0	0				25%	8%	1	0	0	100%	25%	0	0	1				50%	9%
1422	1	1	8	10%	50%	0	0	1				9%	12%	1	1	3	20%	38%	0	0	1				17%	14%
1442	2	0	5	29%	67%							29%	15%	2	0	1	67%	50%							67%	18%
1447	0	2	9	0%	33%	1	1	8				5%	19%	1	0	3	25%	25%	1	0	5				20%	18%
1501	1	0	3	25%	33%	0	0	3	0	1	1	11%	12%	0	1	3	0%	13%							0%	5%
1510	1	0	6	14%	33%	0	0	2				11%	8%	1	0	2	33%	25%	0	1	2				17%	14%
1543	1	0	1	50%	33%	2	0	0	0	1	1	50%	27%	2	0	0	100%	50%	0	0	1	0	0	1	50%	18%
1544	0	0	2	0%	0%	0	0	2	0	0	1	0%	0%	0	0	2	0%	0%	0	0	1	0	0	1	0%	0%
1554	0	0	3	0%	0%	0	1	2	0	0	1	0%	4%	1	0	1	50%	25%	1	0	3	1	0	1	38%	27%
Avgs.	**0.6**	**0.5**	**4.6**	**13%**	**29%**	**0.4**	**0.3**	**2.0**	**0.0**	**0.3**	**1.2**	**14%**	**11%**	**1.2**	**0.3**	**1.9**	**43%**	**33%**	**0.2**	**0.1**	**1.8**	**0.3**	**0.0**	**1.0**	**29%**	**15%**
Experts	3					5			5					4					3			4				
Treatment Group 1st Evaluation, Category 4.0 - Human Resource Development & Management																										
	Item 4.1					Areas for			Site Visit			Item 4.1		Item 4.2					Areas for			Site Visit			Item 4.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2408	1	1	8	10%	50%	0	0	4				7%	12%	2	0	2	50%	50%	0	0	3				29%	18%
2414	0	0	10	0%	0%	0	0	5	0	0	5	0%	0%	1	1	2	25%	38%	0	1	2	0	1	2	10%	23%
2416	0	1	1	0%	17%							0%	4%	0	1	0	0%	13%	1	0	0				50%	14%
2423	1	0	6	14%	33%	0	0	1	0	0	1	11%	8%	1	0	3	25%	25%	0	0	1	0	0	1	17%	9%
2426	0	1	4	0%	17%	0	1	3	0	0	2	0%	8%	1	0	1	50%	25%	0	0	2	0	1	1	17%	14%
2427	0	1	2	0%	17%							0%	4%	0	0	2	0%	0%				1	0	0	33%	9%
2429	0	0	2	0%	0%							0%	0%	0	2	3	0%	25%	0	0	1				0%	9%
2434	1	0	1	50%	33%	2	0	0	0	2	0	50%	31%	1	0	1	50%	25%	0	0	3	0	1	0	17%	14%
2435	1	1	9	9%	50%	3	0	2	0	0	1	24%	35%	2	0	1	67%	50%	0	1	1	0	1	0	33%	27%
2505	0	0	2	0%	0%	1	0	5				13%	8%	0	2	0	0%	25%	0	1	2				0%	14%
2519	1	0	10	9%	33%	0	0	2				8%	8%	2	0	5	29%	50%	0	0	2				22%	18%
2535	0	0	4	0%	0%	0	0	1				0%	0%	0	2	0	0%	25%	0	0	1				0%	9%
2550	0	0	3	0%	0%				1	0	1	20%	8%	0	1	1	0%	13%	0	0	1				0%	5%
2551	0	0	4	0%	0%	0	0	3				0%	0%	0	2	0	0%	25%	1	0	0				33%	18%
Avgs.	**0.4**	**0.4**	**4.7**	**7%**	**18%**	**0.6**	**0.1**	**2.6**	**0.2**	**0.3**	**1.7**	**9%**	**9%**	**0.7**	**0.8**	**1.5**	**21%**	**28%**	**0.2**	**0.2**	**1.5**	**0.2**	**0.7**	**0.7**	**19%**	**14%**
Overall	**0.5**	**0.4**	**4.7**	**10%**	**23%**	**0.5**	**0.2**	**2.3**	**0.1**	**0.3**	**1.4**	**11%**	**10%**	**0.9**	**0.6**	**1.7**	**31%**	**30%**	**0.2**	**0.2**	**1.6**	**0.2**	**0.4**	**0.8**	**23%**	**15%**
Avgs.																										

CLTI-C4.0_Q5CANALYSIS

Count Data from Q5 Content Analysis for Category 4.0

Control																										
	Item 4.3					Areas for			Site Visit			Item 4.3		Item 4.4					Areas for			Site Visit			Item 4.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	0	2	3	0%	20%							0%	7%	1	0	3	25%	25%							25%	8%
1410	2	0	4	33%	40%	0	0	1	0	0	1	25%	14%	1	1	1	33%	38%	0	0	1	0	0	1	20%	13%
1415	1	1	1	33%	30%	0	0	1				25%	11%	1	0	2	33%	25%	0	1	0				25%	13%
1422	2	0	2	50%	40%	0	0	1				40%	14%	3	0	4	43%	75%	0	0	1				38%	25%
1442	1	4	1	17%	60%							17%	21%	3	0	2	60%	75%							60%	25%
1447	2	1	4	29%	50%	0	1	2				20%	21%	2	0	2	50%	50%	1	0	1				50%	25%
1501	1	3	2	17%	50%	0	0	1	0	0	1	13%	18%	3	1	1	60%	88%							60%	29%
1510	0	4	2	0%	40%	0	1	0				0%	18%	2	0	2	50%	50%	0	0	2				33%	17%
1543	1	1	2	25%	30%	1	0	1	0	0	2	25%	18%	2	0	0	100%	50%	1	0	1	1	0	1	67%	33%
1544	0	0	2	0%	0%	0	1	0	0	0	1	0%	4%	2	0	0	100%	50%	0	0	1	0	0	1	50%	17%
1554	0	0	1	0%	0%	1	0	2	1	0	1	33%	14%	0	0	2	0%	0%	1	0	1	0	0	1	20%	8%
Avgs.	0.9	1.5	2.2	19%	33%	0.2	0.3	1.0	0.2	0.0	1.2	18%	15%	1.8	0.2	1.7	50%	48%	0.4	0.1	1.0	0.3	0.0	1.0	41%	19%
Experts	5					3			6					4					4			4				
Treatme																										
	Item 4.3					Areas for			Site Visit			Item 4.3		Item 4.4					Areas for			Site Visit			Item 4.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2408	0	2	5	0%	20%	0	1	5				0%	11%	2	0	4	33%	50%	0	0	2				25%	17%
2414	0	1	5	0%	10%	0	0	3	0	0	3	0%	4%	2	1	2	40%	63%	1	0	3	2	0	2	38%	46%
2416	1	0	1	50%	20%	0	0	1				33%	7%	0	1	0	0%	13%	0	0	1				0%	4%
2423	2	1	1	50%	50%	0	1	0	0	1	0	33%	25%	3	0	3	50%	75%	0	0	1	0	0	1	38%	25%
2426	0	2	2	0%	20%	0	1	1	0	1	0	0%	14%	2	0	1	67%	50%	0	1	1	0	0	2	29%	21%
2427	1	0	0	100%	20%	1	0	0	0	0	1	67%	14%	1	1	1	33%	38%				0	0	1	25%	13%
2429	1	1	0	50%	30%	0	0	1	0	1	0	25%	14%	2	0	1	67%	50%				0	1	0	50%	21%
2434	2	0	1	67%	40%	0	0	2	0	2	0	29%	21%	2	0	0	100%	50%	1	0	2	0	0	1	50%	25%
2435	2	1	2	40%	50%	2	0	2	0	1	0	40%	36%	2	0	3	40%	50%	0	0	1				33%	17%
2505	0	1	4	0%	10%	0	2	1				0%	11%	0	1	2	0%	13%	1	2	1				14%	21%
2519	2	0	6	25%	40%	0	0	1				22%	14%	3	0	4	43%	75%	0	1	0				38%	29%
2535	0	1	3	0%	10%							0%	4%	1	1	1	33%	38%							33%	13%
2550	1	0	1	50%	20%	0	0	1				33%	7%	0	1	2	0%	13%				1	0	0	25%	13%
2551	0	0	1	0%	0%	0	0	1				0%	0%	0	0	2	0%	0%	1	1	0				25%	13%
Avgs.	0.9	0.7	2.3	31%	24%	0.2	0.4	1.5	0.0	0.9	0.6	20%	13%	1.4	0.4	1.9	36%	41%	0.4	0.5	1.2	0.4	0.1	1.0	30%	20%
Overall	0.9	1.0	2.2	25%	28%	0.2	0.4	1.3	0.1	0.5	0.8	19%	14%	1.6	0.3	1.8	42%	44%	0.4	0.3	1.1	0.4	0.1	1.0	35%	20%
Avgs.																										

CL_TL-C4.0_Q5CANAL1.XLS

Count Data from Q5 Content Analysis for Category 5.0

Control Group 1st Evaluation, Category 5.0 - Process Management																										
	Item 5.1					Areas for			Site Visit			Item 5.1		Item 5.2					Areas for			Site Visit			Item 5.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1420	0	2	4	0%	25%	1	1	5				8%	19%	0	1	4	0%	13%	0	2	4				0%	13%
1430	0	1	4	0%	13%	0	0	3				0%	4%	1	0	3	25%	25%	0	1	1				17%	13%
1440	0	0	3	0%	0%	0	1	0	0	0	1	0%	4%	0	0	1	0%	0%	0	0	2	0	0	1	0%	0%
1441	1	1	3	20%	38%							20%	12%	1	0	2	33%	25%	0	0	1				25%	8%
1443	4	0	1	80%	100%	0	0	1	0	0	1	57%	31%	0	1	4	0%	13%	0	1	0				0%	8%
1513	2	0	5	29%	50%	0	0	3	0	0	2	17%	15%	1	1	5	14%	38%	0	1	1	1	0	0	20%	25%
1523	2	0	4	33%	50%	0	0	4	0	0	1	18%	15%	0	0	4	0%	0%	0	1	1	0	0	2	0%	4%
1525	1	0	2	33%	25%	0	1	5	1	0	1	18%	19%	0	1	1	0%	13%	1	0	6	0	0	1	10%	13%
1544	1	0	1	50%	25%	0	0	1	0	0	1	25%	8%	0	1	1	0%	13%	0	0	1	0	0	1	0%	4%
1552	0	1	1	0%	13%	0	0	2	0	0	2	0%	4%	0	1	1	0%	13%	0	2	0	0	0	1	0%	13%
Avgs.	**1.1**	**0.5**	**2.8**	**25%**	**34%**	**0.1**	**0.3**	**2.7**	**0.1**	**0.0**	**1.3**	**16%**	**13%**	**0.3**	**0.6**	**2.6**	**7%**	**15%**	**0.1**	**0.8**	**1.7**	**0.2**	**0.0**	**1.0**	**7%**	**10%**
Experts	4					4			5					4					3			5				

Treatment Group 1st Evaluation, Category 5.0 - Process Management																										
	Item 5.1					Areas for			Site Visit			Item 5.1		Item 5.2					Areas for			Site Visit			Item 5.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2404	0	1	3	0%	13%	0	0	2	0	0	1	0%	4%	0	0	3	0%	0%	0	0	1	0	0	1	0%	0%
2408	1	2	7	10%	50%	1	0	2				15%	23%	1	1	3	20%	38%	1	0	3				22%	21%
2419	0	2	3	0%	25%	0	0	1	1	0	1	13%	15%	0	0	5	0%	0%	1	1	1	0	0	2	10%	13%
2432	0	1	5	0%	13%	0	0	1				0%	4%	0	0	4	0%	0%	0	0	2				0%	0%
2433	1	1	5	14%	38%	0	0	4				9%	12%	0	0	4	0%	0%	0	0	3				0%	0%
2434	1	0	2	33%	25%	0	1	3	2	0	1	30%	27%	0	0	3	0%	0%	0	0	4	0	0	3	0%	0%
2444	1	1	4	17%	38%	0	0	2				13%	12%	0	1	4	0%	13%	0	0	1				0%	4%
2504	0	1	4	0%	13%	0	0	3				0%	4%	0	2	4	0%	25%	0	0	2				0%	8%
2531	2	2	6	20%	75%							20%	23%	0	1	8	0%	13%							0%	4%
2533	0	0	3	0%	0%	1	1	2	0	0	1	13%	12%	0	0	7	0%	0%	0	1	3	0	0	2	0%	4%
2545	3	1	6	30%	88%	0	0	2	0	0	1	23%	27%	1	1	2	25%	38%	0	1	0	0	0	1	17%	17%
2546	3	0	6	33%	75%	0	1	2				25%	27%	2	1	4	29%	63%	0	0	3	0	0	1	18%	21%
Avgs.	**1.0**	**1.0**	**4.5**	**13%**	**38%**	**0.2**	**0.3**	**2.2**	**0.6**	**0.0**	**1.0**	**13%**	**16%**	**0.3**	**0.6**	**4.3**	**6%**	**16%**	**0.2**	**0.3**	**2.1**	**0.0**	**0.0**	**1.7**	**6%**	**8%**
Overall	**1.0**	**0.8**	**3.7**	**18%**	**36%**	**0.2**	**0.3**	**2.4**	**0.3**	**0.0**	**1.2**	**15%**	**15%**	**0.3**	**0.6**	**3.5**	**7%**	**15%**	**0.1**	**0.5**	**1.9**	**0.1**	**0.0**	**1.3**	**6%**	**9%**
Avgs.																										

C1, T1-5.0, Q5CANAL1.XLS

Count Data from Q5 Content Analysis for Category 5.0

Control																										
	Item 5.3					Areas for			Site Visit			Item 5.3		Item 5.4					Areas for			Site Visit			Item 5.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1420	0	1	5	0%	13%	1	1	2				10%	17%	1	1	6	13%	30%	0	0	2				10%	13%
1430	0	1	3	0%	13%	0	0	2				0%	4%	0	1	5	0%	10%	0	0	1				0%	4%
1440	0	1	3	0%	13%	0	0	1	0	0	1	0%	4%	2	0	1	67%	40%	0	0	2				40%	17%
1441	1	1	4	17%	38%							17%	13%	2	0	1	67%	40%	0	0	1				50%	17%
1443	1	0	5	17%	25%	1	2	0	1	0	0	30%	33%	3	2	1	50%	80%	0	0	1				43%	33%
1513	2	0	4	33%	50%	0	1	1	0	0	1	22%	21%	3	0	2	60%	60%	0	0	2	0	0	1	38%	25%
1523	1	1	3	20%	38%	0	1	3	0	0	1	10%	17%	1	1	4	17%	30%	0	0	3				11%	13%
1525	0	1	3	0%	13%	2	1	2	0	1	1	18%	29%	1	1	2	25%	30%	0	0	5	0	0	2	9%	13%
1544	0	0	2	0%	0%	0	0	1	0	0	1	0%	0%	1	0	1	50%	20%	0	0	1	0	0	1	25%	8%
1552	1	0	2	33%	25%	0	0	2	0	0	1	17%	8%	1	0	1	50%	20%	0	0	2	0	1	0	20%	13%
Avgs.	**0.6**	**0.6**	**3.4**	**12%**	**23%**	**0.4**	**0.7**	**1.6**	**0.1**	**0.1**	**0.9**	**12%**	**15%**	**1.5**	**0.6**	**2.4**	**40%**	**36%**	**0.0**	**0.0**	**2.0**	**0.0**	**0.3**	**1.0**	**25%**	**15%**
Experts	4					4			4					5					1			6				
Treatme																										
	Item 5.3					Areas for			Site Visit			Item 5.3		Item 5.4					Areas for			Site Visit			Item 5.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2404	1	0	3	25%	25%							25%	8%	1	0	0	100%	20%	0	0	1				50%	8%
2408	1	1	9	9%	38%	1	0	5				12%	21%	2	0	4	33%	40%	0	0	3				22%	17%
2419	0	0	5	0%	0%	0	2	1	0	0	1	0%	8%	3	0	1	75%	60%	0	0	2	0	1	0	43%	29%
2432	0	1	3	0%	13%	1	0	1				17%	13%	1	1	3	20%	30%	0	0	1				17%	13%
2433	0	2	2	0%	25%	1	1	1				14%	21%	1	1	3	20%	30%	0	0	2				14%	13%
2434	0	1	1	0%	13%	0	0	4	0	0	1	0%	4%	3	0	1	75%	60%	0	0	3	1	0	3	36%	33%
2444	1	0	7	13%	25%	0	2	1				9%	17%	2	1	1	50%	50%	0	0	1				40%	21%
2504	1	1	4	17%	38%	2	1	1				30%	33%	2	1	3	33%	50%	0	0	2				25%	21%
2531	1	1	10	8%	38%	0	0	1				8%	13%	4	0	6	40%	80%	0	0	2				33%	33%
2533	1	0	5	17%	25%	0	0	4	0	0	2	8%	8%	2	0	1	67%	40%	0	0	2	1	0	0	50%	25%
2545	2	1	6	22%	63%	0	1	0	0	0	1	18%	25%	1	0	5	17%	20%	0	0	1	0	0	1	13%	8%
2546	0	2	7	0%	25%	0	1	1				0%	13%	2	0	7	22%	40%	0	0	1				20%	17%
Avgs.	**0.7**	**0.8**	**5.2**	**9%**	**27%**	**0.5**	**0.7**	**1.8**	**0.0**	**0.0**	**1.3**	**12%**	**15%**	**2.0**	**0.3**	**2.9**	**46%**	**43%**	**0.0**	**0.0**	**1.8**	**0.5**	**0.3**	**1.0**	**30%**	**20%**
Overall	**0.6**	**0.7**	**4.4**	**10%**	**25%**	**0.5**	**0.7**	**1.7**	**0.1**	**0.1**	**1.0**	**12%**	**15%**	**1.8**	**0.5**	**2.7**	**43%**	**40%**	**0.0**	**0.0**	**1.9**	**0.3**	**0.3**	**1.0**	**28%**	**18%**
Avgs.																										

C1, T1-5.0, Q5CANAL1.XLS

Count Data from Q5 Content Analysis for Category 6.0

Control Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.1					Areas for			Site Visit			Item 6.1	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1428	1	0	0	100%	25%	0	0	1				50%	8%
1431	2	0	0	100%	50%	0	0	2				50%	17%
1503	1	1	0	50%	38%							50%	13%
1510	2	0	0	100%	50%	1	0	0				100%	25%
1528	2	0	0	100%	50%	0	0	3				40%	17%
1530	1	0	0	100%	25%	0	0	2				33%	8%
1534	2	0	0	100%	50%	0	0	1	0	0	2	40%	17%
Avgs.	**1.6**	**0.1**	**0.0**	**93%**	**41%**	**0.2**	**0.0**	**1.5**	**0.0**	**0.0**	**2.0**	**52%**	**15%**
Experts	4					4			4				
Treatment Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.1					Areas for			Site Visit			Item 6.1	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2401	1	0	0	100%	25%	0	0	1	0	0	1	33%	8%
2402	0	1	1	0%	13%	1	0	3				17%	13%
2404	1	0	0	100%	25%	1	0	0				100%	17%
2405	2	0	0	100%	50%	0	0	5	0	0	2	22%	17%
2412	3	0	0	100%	75%	0	0	1	0	0	1	60%	25%
2426	1	1	1	33%	38%	1	0	1	0	0	2	29%	21%
2439	1	2	0	33%	50%	0	0	1	0	0	1	20%	17%
2507	2	0	0	100%	50%	1	0	0				100%	25%
2509	3	0	0	100%	75%	0	1	0	0	0	1	60%	29%
2519	1	1	0	50%	38%	0	0	2				25%	13%
2535	1	1	0	50%	38%							50%	13%
2536	1	0	0	100%	25%	0	0	1				50%	8%
Avgs.	**1.4**	**0.5**	**0.2**	**72%**	**42%**	**0.4**	**0.1**	**1.4**	**0.0**	**0.0**	**1.3**	**47%**	**17%**
Overall	**1.5**	**0.4**	**0.1**	**80%**	**41%**	**0.3**	**0.1**	**1.4**	**0.0**	**0.0**	**1.4**	**49%**	**16%**
Avgs.													

Control Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.2					Areas for			Site Visit			Item 6.2	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1428					0%	0	1	0	1	0	0	50%	14%
1431	0	1	0	0%	17%	0	0	4				0%	5%
1503	0	1	0	0%	17%							0%	5%
1510	1	0	0	100%	33%	1	0	1				67%	18%
1528	2	0	0	100%	67%	0	0	8				20%	18%
1530	2	0	0	100%	67%	0	0	2				50%	18%
1534	1	1	0	50%	50%	0	0	2	0	0	4	13%	14%
Avgs.	**1.0**	**0.5**	**0.0**	**58%**	**36%**	**0.2**	**0.2**	**2.8**	**0.5**	**0.0**	**2.0**	**28%**	**13%**
Experts	3					4			4				
Treatment Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.2					Areas for			Site Visit			Item 6.2	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2401	2	1	0	67%	83%	1	0	0	0	1	0	60%	36%
2402	2	0	0	100%	67%	0	1	4				29%	23%
2404	0	1	0	0%	17%	1	0	0				50%	14%
2405	2	1	0	67%	83%	0	0	5				25%	23%
2412	3	0	0	100%	100%	1	0	3	0	0	1	50%	36%
2426	0	1	2	0%	17%	0	0	2	0	1	1	0%	9%
2439	1	0	0	100%	33%	0	0	1	0	0	1	33%	9%
2507	0	2	0	0%	33%	0	0	2				0%	9%
2509	1	2	0	33%	67%	1	0	2	0	1	1	25%	32%
2519	1	0	0	100%	33%	1	0	3				40%	18%
2535	1	1	0	50%	50%	0	0	3				20%	14%
2536	0	1	0	0%	17%	0	0	1				0%	5%
Avgs.	**1.1**	**0.8**	**0.2**	**51%**	**50%**	**0.4**	**0.1**	**2.2**	**0.0**	**0.6**	**0.8**	**28%**	**19%**
Overall	**1.1**	**0.7**	**0.1**	**54%**	**45%**	**0.3**	**0.1**	**2.4**	**0.1**	**0.4**	**1.1**	**28%**	**17%**
Avgs.													

Control Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.3					Areas for			Site Visit			Item 6.3	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1428	0	1	0	0%	10%							0%	4%
1431	1	0	0	100%	20%	0	0	1				50%	8%
1503	0	1	0	0%	10%							0%	4%
1510	1	1	0	50%	30%	1	0	0				67%	21%
1528	2	1	0	67%	50%	1	0	2				50%	29%
1530	0	2	1	0%	20%	0	0	1				0%	8%
1534	0	1	1	0%	10%	0	1	1	0	0	1	0%	8%
Avgs.	**0.6**	**1.0**	**0.3**	**31%**	**21%**	**0.4**	**0.2**	**1.0**	**0.0**	**0.0**	**1.0**	**24%**	**12%**
Experts	5					3			4				
Treatment Group 1st Evaluation, Category 6.0 - Business Results													
	Item 6.3					Areas for			Site Visit			Item 6.3	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2401	2	2	0	50%	60%	0	0	2	0	1	0	29%	29%
2402	1	0	2	33%	20%	0	0	3				17%	8%
2404	1	1	1	33%	30%							33%	13%
2405	1	2	1	25%	40%	1	0	3	0	0	1	22%	25%
2412	2	1	4	29%	50%	0	0	1	0	0	1	22%	21%
2426	2	0	0	100%	40%	0	0	2	0	1	0	40%	21%
2439	0	0	2	0%	0%	0	1	1	0	0	1	0%	4%
2507	3	0	2	60%	60%	0	0	1				50%	25%
2509	0	0	3	0%	0%	0	1	2	0	0	1	0%	4%
2519	2	2	1	40%	60%	0	0	1				33%	25%
2535	1	0	2	33%	20%							33%	8%
2536	0	0	1	0%	0%	0	0	1				0%	0%
Avgs.	**1.3**	**0.7**	**1.6**	**34%**	**32%**	**0.1**	**0.2**	**1.7**	**0.0**	**0.3**	**0.7**	**23%**	**15%**
Overall	**1.0**	**0.8**	**1.1**	**33%**	**28%**	**0.2**	**0.2**	**1.5**	**0.0**	**0.3**	**0.7**	**23%**	**14%**
Avgs.													

Count Data from Q5 Content Analysis for Category 7.0

Control Group 1st Evaluation, Category 7.0 - Customer Focus & Satisfaction

	Item 7.1					Areas for			Site Visit			Item 7.1		Item 7.2					Areas for			Site Visit			Item 7.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	1	2	2	20%	40%							20%	17%	4	0	1	80%	80%							80%	31%
1417	1	2	2	20%	40%							20%	17%	4	1	1	67%	90%							67%	35%
1420	0	1	2	0%	10%	1	0	3				14%	13%	1	0	2	33%	20%	0	0	2				20%	8%
1422	3	0	5	38%	60%	0	0	1				33%	25%	3	1	4	38%	70%	0	0	1				33%	27%
1428	0	1	0	0%	10%							0%	4%	0	1	1	0%	10%							0%	4%
1430	3	1	1	60%	70%	0	0	1				50%	29%	3	1	2	50%	70%	0	0	1				43%	27%
1440	0	2	1	0%	20%	0	1	1				0%	13%	0	1	0	0%	10%	0	0	1	0	0	2	0%	4%
1447	1	1	7	11%	30%	1	1	3				14%	25%	1	2	8	9%	40%	0	0	3				7%	15%
1522	1	0	0	100%	20%	0	0	2				33%	8%	0	0	2	0%	0%							0%	0%
1527	0	2	3	0%	20%	0	0	2	0	0	1	0%	8%	0	1	3	0%	10%	0	1	2	0	0	1	0%	8%
1534	2	1	2	40%	50%				1	0	1	43%	29%	1	1	4	17%	30%	0	0	3	0	0	2	9%	12%
1543	2	0	2	50%	40%	0	0	2	0	0	2	25%	17%	3	1	1	60%	70%	0	0	1	0	0	1	43%	27%
Avgs.	**1.2**	**1.1**	**2.3**	**28%**	**34%**	**0.3**	**0.3**	**1.9**	**0.3**	**0.0**	**1.3**	**21%**	**17%**	**1.7**	**0.8**	**2.4**	**29%**	**42%**	**0.0**	**0.1**	**1.8**	**0.0**	**0.0**	**1.5**	**25%**	**16%**
Experts	5					3			4					5					3			5				

Treatment Group 1st Evaluation, Category 7.0 - Customer Focus & Satisfaction

	Item 7.1					Areas for			Site Visit			Item 7.1		Item 7.2					Areas for			Site Visit			Item 7.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2407	1	2	2	20%	40%							20%	17%	3	0	3	50%	60%							50%	23%
2412	3	1	2	50%	70%	1	0	1	0	0	1	44%	38%	3	2	4	33%	80%	0	0	1	0	0	1	27%	31%
2435	1	0	1	50%	20%							50%	8%	0	3	1	0%	30%	0	0	1	0	0	1	0%	12%
2444	2	1	4	29%	50%	0	1	3				18%	25%	0	2	5	0%	20%	0	0	6				0%	8%
2509	1	0	3	25%	20%	1	0	0				40%	17%	2	1	5	25%	50%	0	0	1	0	0	2	18%	19%
2515	1	3	3	14%	50%	1	0	3	0	0	3	14%	29%	1	2	3	17%	40%	1	0	3	0	0	2	17%	23%
2521	1	1	0	50%	30%	0	0	2				25%	13%	0	1	2	0%	10%							0%	4%
2532	1	1	2	25%	30%	0	0	3				14%	13%	0	2	3	0%	20%	0	0	3				0%	8%
2533	1	2	1	25%	40%	0	0	2	0	0	1	14%	17%	0	1	5	0%	10%							0%	4%
2548	1	1	7	11%	30%	1	0	3				15%	21%	3	1	5	33%	70%	1	0	2				33%	35%
2550	0	1	4	0%	10%	0	0	1	0	0	1	0%	4%	0	2	4	0%	20%	0	0	1	0	0	2	0%	8%
2553	2	0	9	18%	40%	0	0	2				15%	17%	2	2	3	29%	60%							29%	23%
Avgs.	**1.3**	**1.1**	**3.2**	**26%**	**36%**	**0.4**	**0.1**	**2.0**	**0.0**	**0.0**	**1.5**	**23%**	**18%**	**1.2**	**1.6**	**3.6**	**16%**	**39%**	**0.3**	**0.0**	**2.3**	**0.0**	**0.0**	**1.6**	**15%**	**16%**
Overall **Avgs.**	**1.2**	**1.1**	**2.7**	**27%**	**35%**	**0.3**	**0.2**	**1.9**	**0.1**	**0.0**	**1.4**	**22%**	**18%**	**1.4**	**1.2**	**3.0**	**23%**	**40%**	**0.1**	**0.1**	**2.0**	**0.0**	**0.0**	**1.6**	**20%**	**16%**

Count Data from Q5 Content Analysis for Category 7.0

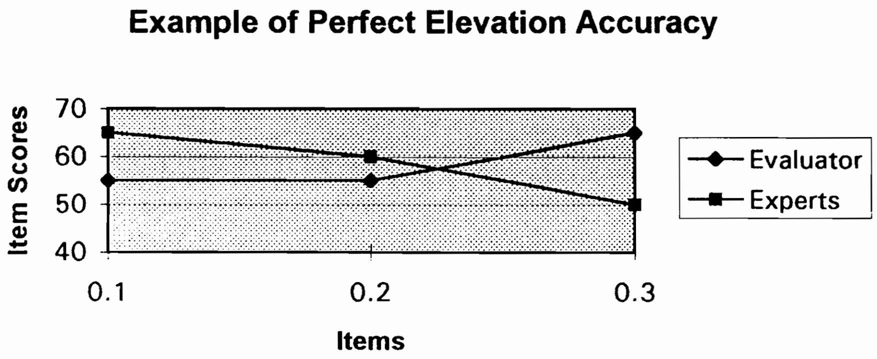
Control																																							
	Item 7.3					Areas for			Site Visit			Item 7.3		Item 7.4					Areas for			Site Visit			Item 7.4		Item 7.5					Areas for			Site Visit			Item 7.5	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Ex
1403	2	2	0	50%	60%							50%	20%	1	0	0	100%	20%	0	0	1				50%	9%	0	1	1	0%	17%							0%	4%
1417	1	0	1	50%	20%							50%	7%	0	2	0	0%	20%							0%	9%	1	0	0	100%	33%							100%	8%
1420	0	0	1	0%	0%	1	0	2				25%	7%	0	2	1	0%	20%	0	0	1				0%	9%					0%	1	0	2				33%	8%
1422	1	2	2	20%	40%	1	0	0				33%	20%	2	1	3	33%	50%	0	0	1				29%	23%	1	0	1	50%	33%							50%	8%
1428	0	1	0	0%	10%							0%	3%	0	1	0	0%	10%							0%	5%					0%								0%
1430	0	1	3	0%	10%	0	0	1				0%	3%	0	0	4	0%	0%	0	0	1				0%	0%	0	0	2	0%	0%	0	0	1				0%	0%
1440	0	1	0	0%	10%	0	0	2				0%	3%	1	0	2	33%	20%	0	0	1	0	0	2	17%	9%	0	0	3	0%	0%	0	0	2	0	1	2	0%	4%
1447	0	2	5	0%	20%	1	1	3				8%	17%	0	0	1	0%	0%	0	0	1				0%	0%	0	0	2	0%	0%	0	0	1				0%	0%
1522	0	1	1	0%	10%	1	0	0				33%	10%	0	0	2	0%	0%							0%	0%	0	0	3	0%	0%	0	0	1				0%	0%
1527	0	0	2	0%	0%	0	0	1				0%	0%	0	0	2	0%	0%	0	0	2	0	0	1	0%	0%	0	0	2	0%	0%	0	0	2				0%	0%
1534	1	1	1	33%	30%	0	0	1	0	0	1	20%	10%	1	0	2	33%	20%	0	1	1	0	0	1	17%	14%	1	0	1	50%	33%	0	0	1				33%	8%
1543	3	1	1	60%	70%							60%	23%	2	0	0	100%	40%	0	0	1	0	0	1	50%	18%	1	0	1	50%	33%							50%	8%
Avgs.	**0.7**	**1.0**	**1.4**	**18%**	**23%**	**0.5**	**0.1**	**1.3**	**0.0**	**0.0**	**1.0**	**23%**	**10%**	**0.6**	**0.5**	**1.4**	**25%**	**17%**	**0.0**	**0.1**	**1.1**	**0.0**	**0.0**	**1.3**	**13%**	**8%**	**0.4**	**0.1**	**1.6**	**25%**	**13%**	**0.1**	**0.0**	**1.4**	**0.0**	**1.0**	**2.0**	**24%**	**4%**
Experts	5					6			4					5					4			2					3					5			4				
Treatme																																							
	Item 7.3					Areas for			Site Visit			Item 7.3		Item 7.4					Areas for			Site Visit			Item 7.4		Item 7.5					Areas for			Site Visit			Item 7.5	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Ex
2407	4	0	0	100%	80%	0	1	1	0	1	1	50%	33%	1	0	1	50%	20%	0	0	1	0	0	1	25%	9%	1	0	1	50%	33%	0	0	1	0	1	0	25%	13%
2412	2	2	0	50%	60%	1	0	3				38%	27%	3	0	0	100%	60%	1	0	1				80%	36%	1	0	1	50%	33%				0	0	1	33%	8%
2435	1	2	0	33%	40%	0	0	1				25%	13%	3	1	1	60%	70%	0	0	1				50%	32%	1	0	1	50%	33%							50%	8%
2444	1	1	5	14%	30%	0	1	1				11%	13%	0	1	2	0%	10%	1	0	2				17%	14%	0	0	3	0%	0%	0	0	4				0%	0%
2509	2	0	1	67%	40%	0	1	2	1	0	0	43%	23%	0	1	1	0%	10%	0	0	2	0	1	0	0%	9%	0	0	2	0%	0%	0	0	2	0	1	0	0%	4%
2515	1	1	3	20%	30%	0	0	3	0	0	1	11%	10%	2	2	1	40%	60%	1	0	2	0	1	2	27%	41%	1	1	1	33%	50%	0	2	0	0	2	0	14%	29%
2521	0	1	1	0%	10%	0	0	1				0%	3%	0	1	1	0%	10%							0%	5%	0	0	2	0%	0%							0%	0%
2532	1	0	1	50%	20%	0	0	1				33%	7%	0	2	0	0%	20%	0	0	1				0%	9%	0	0	1	0%	0%	0	0	2				0%	0%
2533	2	1	2	40%	50%	0	0	3	0	0	2	20%	17%	0	1	1	0%	10%	0	0	3				0%	5%	0	0	1	0%	0%	0	2	2	0	1	0	0%	13%
2548	3	0	1	75%	60%	0	0	2				50%	20%	4	0	0	100%	80%	0	1	0				80%	41%	1	0	1	50%	33%	1	1	4				25%	21%
2550	1	0	0	100%	20%				0	0	2	33%	7%	1	0	0	100%	20%							100%	9%	0	0	1	0%	0%							0%	0%
2553	1	2	2	20%	40%	0	0	3				13%	13%	1	2	1	25%	40%	0	0	1				20%	18%	0	0	2	0%	0%							0%	0%
Avgs.	**1.6**	**0.8**	**1.3**	**47%**	**40%**	**0.1**	**0.3**	**1.9**	**0.2**	**0.2**	**1.2**	**27%**	**16%**	**1.3**	**0.9**	**0.8**	**40%**	**34%**	**0.3**	**0.1**	**1.4**	**0.0**	**0.7**	**1.0**	**33%**	**19%**	**0.4**	**0.1**	**1.4**	**19%**	**15%**	**0.1**	**0.7**	**2.1**	**0.0**	**1.0**	**0.2**	**12%**	**8%**
Overall	**1.1**	**0.9**	**1.4**	**33%**	**32%**	**0.3**	**0.2**	**1.6**	**0.2**	**0.2**	**1.2**	**25%**	**13%**	**0.9**	**0.7**	**1.1**	**32%**	**25%**	**0.2**	**0.1**	**1.3**	**0**	**0.3**	**1.1**	**23%**	**13%**	**0.4**	**0.1**	**1.5**	**22%**	**14%**	**0.1**	**0.4**	**1.8**	**0**	**1.0**	**0.5**	**18%**	**6%**
Avgs.																																							

Appendix AT. Elevation and Dimensional Accuracy Illustrated

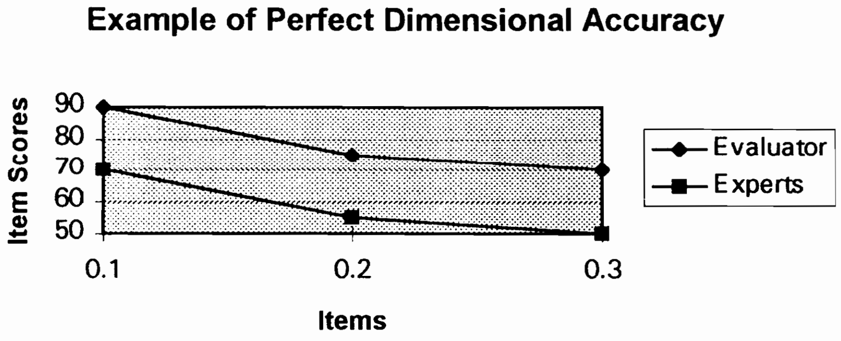
Elevation (E) = $\sqrt{(\bar{x}.-\bar{t}.)^2}$

Dimensional Accuracy (DA) = $\sqrt{\frac{1}{n}\sum\left[\left(x_j-\bar{x}.\right)-\left(t_j-\bar{t}.\right)\right]^2}$

The following charts illustrate the concepts of Dimensional Accuracy (DA) and Elevation Accuracy (E). Please note these examples were oversimplified for illustrative purposes.



In the above chart, the Evaluator's mean score is 58.3, the same as the Experts' mean score. While the Evaluator's Elevation Accuracy is perfect (E = 0), the Evaluator's Dimensional Accuracy is poor (DA = 10.8).



In the above chart, the Evaluator's scores and the Experts' scores have a correlation of 1.0 and equal variances. While the Evaluator's Dimensional Accuracy is perfect (DA = 0), the Evaluator's Elevation Accuracy is poor (E = 20).

Appendix AU. Count Data from Q10 Content Analysis

Control Group 2nd Evaluation, Category 1.0 - Leadership

	Item 1.1					Areas for			Site Visit			Item 1.1		Item 1.2					Areas for			Site Visit			Item 1.2		Item 1.3					Areas for			Site Visit			Item 1.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1410	1	3	0	25%	28%	1	0	0	0	0	1	33%	19%	0	4	0	0%	40%				0	0	1	0%	13%	2	1	1	50%	50%	0	1	0	0	0	1	33%	19%
1420	1	3	1	20%	28%	1	2	2				20%	25%	0	1	1	0%	10%	2	1	0				40%	19%	1	1	3	20%	30%	0	2	2				11%	16%
1421	0	1	1	0%	6%	0	0	1				0%	3%	0	2	2	0%	20%	0	0	1				0%	6%	1	2	2	20%	40%	0	0	1				17%	13%
1428	0	0	1	0%	0%							0%	0%	0	1	0	0%	10%							0%	3%					0%								0%
1430	4	0	3	57%	44%	0	0	3				40%	22%	3	0	0	100%	60%	0	0	1				75%	19%	2	1	2	40%	50%	0	0	1				33%	16%
1431	1	3	1	20%	28%							20%	14%	1	2	0	33%	40%							33%	13%	0	0	1	0%	0%	0	1	1				0%	3%
1440	2	1	1	50%	28%	0	0	1	1	0	1	43%	19%	1	1	1	33%	30%							33%	9%	1	1	0	50%	30%	0	1	2	0	0	1		
1442	4	0	1	80%	44%	0	0	1				67%	22%	3	1	0	75%	70%	0	0	1				60%	22%	4	0	2	67%	80%	0	1	0				57%	28%
1543	5	0	4	56%	56%							56%	28%	3	1	0	75%	70%	1	1	0	2	0	0	75%	44%	4	1	1	67%	90%							67%	28%
1544	1	0	1	50%	11%	0	0	1				33%	6%	1	0	0	100%	20%	0	0	1	0	1	0	33%	9%	1	0	1	50%	20%	0	0	1	0	1	0	25%	9%
1552	1	2	0	33%	22%	0	0	1	0	0	1	20%	11%	1	1	0	50%	30%	0	0	1	0	0	1	25%	9%	1	0	1	50%	20%	0	1	0				33%	9%
Avgs.	**1.8**	**1.2**	**1.3**	**36%**	**27%**	**0.3**	**0.3**	**1.3**	**0.3**	**0**	**1**	**30%**	**15%**	**1.2**	**1.3**	**0.4**	**42%**	**36%**	**0.4**	**0.3**	**0.7**	**0.5**	**0.3**	**0.5**	**34%**	**15%**	**1.7**	**0.7**	**1.4**	**41%**	**37%**	**0**	**0.8**	**0.9**	**0**	**0.3**	**0.7**	**31%**	**14%**
Experts	**9**					**4**			**5**					**5**					**6**			**5**					**5**					**6**			**5**				

Treatment Group 2nd Evaluation, Category 1.0 - Leadership

	Item 1.1					Areas for			Site Visit			Item 1.1		Item 1.2					Areas for			Site Visit			Item 1.2		Item 1.3					Areas for			Site Visit			Item 1.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2402	2	1	3	33%	28%							33%	14%	3	1	3	43%	70%							43%	22%	3	1	2	50%	70%	0	1	1				38%	25%
2404	2	2	1	40%	33%							40%	17%	1	0	0	100%	20%	1	0	0				100%	13%	0	1	2	0%	10%	0	1	0	1	0	0	20%	13%
2407	2	2	1	40%	33%				1	0	1	43%	22%	2	2	0	50%	60%							50%	19%	2	1	2	40%	50%				0	0	1	33%	16%
2408	2	2	1	40%	33%	0	0	1	1	0	0	43%	22%	1	2	0	33%	40%	0	0	1	0	1	0	20%	16%	3	2	1	50%	80%				0	0	1	43%	25%
2412	2	0	1	67%	22%	1	0	0	0	0	1	60%	17%	2	1	0	67%	50%	0	0	1	0	1	0	40%	19%	1	1	2	25%	30%	1	1	0				33%	19%
2416	2	1	0	67%	28%	2	0	1	0	0	1	57%	25%	1	2	0	33%	40%	0	2	0	1	0	0	33%	25%	1	1	1	33%	30%	0	3	1	0	1	0	13%	22%
2419	2	2	0	50%	33%	1	0	1				50%	22%	2	1	0	67%	50%	1	1	1				50%	25%	0	4	0	0%	40%	0	1	2				0%	16%
2435	4	0	0	100%	44%	0	0	1	0	1	0	67%	25%	2	0	0	100%	40%	0	0	1	0	1	0	50%	16%	5	0	0	100%	100%							100%	31%
2509	4	0	2	67%	44%				0	0	1	57%	22%	2	2	0	50%	60%	0	0	2	0	1	0	29%	22%	3	1	3	43%	70%							43%	22%
2519	2	2	1	40%	33%	0	0	1	0	1	1	25%	19%	3	0	0	100%	60%	0	1	0	0	1	0	60%	25%	3	0	1	75%	60%	0	0	2	0	0	1	43%	19%
2521					0%	0	1	2	0	1	0	0%	6%	1	0	0	100%	20%	0	0	1				50%	6%	1	0	0	100%	20%	0	1	1				33%	9%
Avgs.	**2.4**	**1.2**	**1**	**54%**	**30%**	**0.6**	**0.1**	**1**	**0.3**	**0.4**	**0.6**	**43%**	**19%**	**1.8**	**1**	**0.3**	**68%**	**46%**	**0.2**	**0.4**	**0.8**	**0.2**	**0.8**	**0**	**48%**	**19%**	**2**	**1.1**	**1.3**	**47%**	**51%**	**0.1**	**1.1**	**1**	**0.2**	**0.2**	**0.6**	**36%**	**20%**
Experts	**9**					**4**			**5**					**5**					**6**			**5**					**5**					**6**			**5**				
Overall																																							
Avgs.	**2.1**	**1.2**	**1.1**	**44%**	**29%**	**0.4**	**0.2**	**1.1**	**0.3**	**0.3**	**0.7**	**37%**	**17%**	**1.5**	**1.1**	**0.3**	**55%**	**41%**	**0.3**	**0.4**	**0.8**	**0.3**	**0.6**	**0.2**	**41%**	**17%**	**1.9**	**0.9**	**1.3**	**44%**	**44%**	**0.1**	**0.9**	**0.9**	**0.1**	**0.3**	**0.6**	**34%**	**17%**

Count Data from Q10 Content Analysis for Category 2.0

Control Group 2nd Evaluation, Category 2.0 - Information & Analysis																																							
	Item 2.1					Areas for			Site Visit			Item 2.1		Item 2.2					Areas for			Site Visit			Item 2.2		Item 2.3					Areas for			Site Visit			Item 2.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1415	1	0	2	33%	20%	0	0	1				25%	8%	1	0	1	50%	33%	0	1	0				33%	17%	1	0	1	50%	25%	0	0	1				33%	9%
1443	2	1	3	33%	50%	1	0	0				43%	29%	2	0	1	67%	67%	1	1	1				50%	39%	1	0	1	50%	25%	1	0	1				50%	18%
1502	0	0	2	0%	0%	0	0	4	0	0	1	0%	0%	0	1	0	0%	17%	0	0	1				0%	6%	0	1	3	0%	13%							0%	5%
1510	1	1	3	20%	30%	1	1	1				25%	25%	1	1	3	20%	50%	0	0	2				14%	17%	1	2	1	25%	50%	1	0	2				29%	27%
1522	1	2	1	25%	40%	0	0	2				17%	17%	0	0	2	0%	0%	1	0	2				20%	11%	0	0	2	0%	0%	0	1	1				0%	5%
1525	2	1	4	29%	50%	1	1	1	0	0	1	27%	33%	1	1	3	20%	50%	1	0	0	1	0	1	38%	39%	0	1	1	0%	13%	2	0	1	1	0	2	38%	32%
1554	0	1	1	0%	10%	1	1	2	0	0	1	14%	17%	0	0	2	0%	0%	0	0	1	0	0	1	0%	0%	0	0	1	0%	0%	0	1	1	0	1	0	0%	9%
Avgs.	**1.0**	**0.9**	**2.3**	**20%**	**29%**	**0.6**	**0.4**	**1.6**	**0.0**	**0.0**	**1.0**	**22%**	**18%**	**0.7**	**0.4**	**1.7**	**22%**	**31%**	**0.4**	**0.3**	**1.0**	**0.5**	**0.0**	**1.0**	**22%**	**18%**	**0.4**	**0.6**	**1.4**	**18%**	**18%**	**0.7**	**0.3**	**1.2**	**0.5**	**0.5**	**1.0**	**21%**	**15%**
Experts	**5**					**4**			**3**					**3**					**3**			**3**					**4**					**3**			**4**				

Treatment Group 2nd Evaluation, Category 2.0 - Information & Analysis																																							
	Item 2.1					Areas for			Site Visit			Item 2.1		Item 2.2					Areas for			Site Visit			Item 2.2		Item 2.3					Areas for			Site Visit			Item 2.3	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2416	1	0	3	25%	20%	1	0	1				33%	17%	1	0	3	25%	33%	0	1	2	0	0	1	13%	17%	0	0	2	0%	0%	0	1	3	0	1	0	0%	9%
2427					0%	1	0	2				33%	8%	0	0	1	0%	0%	0	0	1				0%	0%	0	2	0	0%	25%							0%	9%
2505	1	2	4	14%	40%	0	1	5	0	1	1	7%	25%	0	2	2	0%	33%	0	0	4	0	1	1	0%	17%	1	0	0	100%	25%	0	1	5	0	0	1	13%	14%
2515	0	3	3	0%	30%	0	0	2	0	0	2	0%	13%	0	2	4	0%	33%	1	0	2	0	0	3	8%	22%	0	1	4	0%	13%	1	1	1	0	0	3	9%	18%
2519	4	0	3	57%	80%	0	1	0	0	0	1	44%	38%	3	0	1	75%	100%	0	0	1	0	1	0	50%	39%	3	0	0	100%	75%	0	0	1	0	1	0	60%	32%
2521	1	0	1	50%	20%	0	0	2				25%	8%	0	0	1	0%	0%	0	0	1	0	0	1	0%	0%	0	1	1	0%	13%	1	0	1				25%	14%
2533	0	0	3	0%	0%	0	0	1	0	0	1	0%	0%	0	1	2	0%	17%	0	1	1	1	0	0	17%	22%	1	1	2	25%	38%	1	1	1	0	1	1	22%	32%
2536	2	1	5	25%	50%	0	0	1	0	0	1	20%	21%	1	0	3	25%	33%	0	0	1	0	0	1	17%	11%	1	0	3	25%	25%	0	1	3	0	0	1	11%	14%
2537	0	2	2	0%	20%	0	0	2	0	0	1	0%	8%	0	0	2	0%	0%	0	0	4				0%	0%	1	1	1	33%	38%	1	0	1				40%	23%
2553	3	0	2	60%	60%							60%	25%	0	0	3	0%	0%							0%	0%	0	1	3	0%	13%							0%	5%
Avgs.	**1.3**	**0.9**	**2.9**	**26%**	**32%**	**0.2**	**0.2**	**1.8**	**0.0**	**0.2**	**1.2**	**22%**	**16%**	**0.5**	**0.5**	**2.2**	**13%**	**25%**	**0.1**	**0.2**	**1.9**	**0.1**	**0.3**	**1.0**	**10%**	**13%**	**0.7**	**0.7**	**1.6**	**28%**	**26%**	**0.5**	**0.6**	**2.0**	**0.0**	**0.5**	**1.0**	**18%**	**17%**
Experts	**5**					**4**			**3**					**3**					**3**			**3**					**4**					**3**			**4**				
Overall	**1.2**	**0.9**	**2.6**	**23%**	**31%**	**0.4**	**0.3**	**1.7**	**0.0**	**0.1**	**1.1**	**22%**	**17%**	**0.6**	**0.5**	**2.0**	**17%**	**27%**	**0.3**	**0.3**	**1.5**	**0.2**	**0.2**	**1.0**	**15%**	**15%**	**0.6**	**0.6**	**1.5**	**24%**	**23%**	**0.6**	**0.5**	**1.6**	**0.1**	**0.5**	**1.0**	**19%**	**16%**
Avgs.																																							

Count Data from Q10 Content Analysis for Category 3.0

Control Group 2nd Evaluation, Category 3.0 - Strategic Planning																										
	Item 3.1					Areas for			Site Visit			Item 3.1		Item 3.2					Areas for			Site Visit			Item 3.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	1	1	0	50%	38%	1	0	3	0	0	1	29%	16%	0	2	2	0%	25%	0	0	1	0	0	1	0%	8%
1417	1	0	0	100%	25%	0	1	1	0	0	1	25%	9%	1	0	0	100%	25%	1	0	1	0	0	1	50%	17%
1422	2	1	7	20%	63%	1	0	1				25%	22%	1	1	2	25%	38%	1	1	0				33%	25%
1428	0	1	0	0%	13%				0	0	1	0%	3%					0%								0%
1502	0	0	3	0%	0%	0	0	2				0%	0%	0	1	5	0%	13%							0%	4%
1506	1	0	2	33%	25%	0	1	2	0	0	1	14%	9%	1	0	1	50%	25%	1	0	0	0	1	0	50%	21%
1510	2	1	2	40%	63%	1	0	3				33%	22%	2	1	1	50%	63%	2	1	0				57%	42%
1530	1	2	0	33%	50%	0	0	2				20%	13%	1	1	3	20%	38%							20%	13%
1534	0	3	5	0%	38%	0	0	1				0%	9%	0	1	2	0%	13%	1	0	0				25%	13%
1555	1	1	4	17%	38%	2	1	5	0	0	2	19%	25%	0	2	2	0%	25%	2	2	1	1	1	0	27%	46%
Avgs.	0.9	1.0	2.3	29%	35%	0.6	0.3	2.2	0.0	0.0	1.2	16%	13%	0.7	1.0	2.0	27%	26%	1.1	0.6	0.4	0.3	0.5	0.5	29%	19%
Experts	4					6			6					4					4			4				
Treatment Group 2nd Evaluation, Category 3.0 - Strategic Planning																										
	Item 3.1					Areas for			Site Visit			Item 3.1		Item 3.2					Areas for			Site Visit			Item 3.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2404	2	0	0	100%	50%							100%	13%	0	1	0	0%	13%							0%	4%
2405	1	2	2	20%	50%	0	0	5	0	1	0	9%	16%	0	0	2	0%	0%	0	1	2	0	0	1	0%	4%
2411	2	1	0	67%	63%	0	0	2	0	0	1	33%	16%	1	0	2	33%	25%				0	0	1	25%	8%
2449	1	0	1	50%	25%	1	1	1	0	2	1	25%	22%	0	2	0	0%	25%	3	0	1	0	0	1	43%	33%
2507	2	0	2	50%	50%	1	0	0				60%	19%	0	2	3	0%	25%	1	1	2				11%	21%
2508	2	1	2	40%	63%	0	0	2	0	1	0	25%	19%	2	1	0	67%	63%	0	1	1	0	0	1	33%	25%
2515	1	1	2	25%	38%	0	0	3	0	1	3	9%	13%	1	0	1	50%	25%	2	1	1	0	1	1	38%	33%
2535	0	0	2	0%	0%	0	1	3				0%	3%	0	1	1	0%	13%	0	0	1				0%	4%
2550	1	0	2	33%	25%	0	0	1	1	0	0	40%	13%	0	0	1	0%	0%							0%	0%
Avgs.	1.3	0.6	1.4	43%	40%	0.3	0.3	2.1	0.2	0.8	0.8	34%	15%	0.4	0.8	1.1	17%	21%	1.0	0.7	1.3	0.0	0.2	1.0	17%	15%
Experts	4					6			6					4					4			4				
Overall	1.1	0.8	1.9	36%	38%	0.4	0.3	2.2	0.1	0.5	1.0	25%	14%	0.6	0.9	1.6	22%	24%	1.1	0.6	0.8	0.1	0.3	0.8	23%	17%
Avgs.																										

Q10CANAL.XLS, C2, T2-3.0

Count Data from Q10 Content Analysis for Category 4.0

Control Group 2nd Evaluation, Category 4.0 - Human Resource Development & Management																										
	Item 4.1					Areas for			Site Visit			Item 4.1		Item 4.2					Areas for			Site Visit			Item 4.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1413	0	0	2	0%	0%							0%	0%	0	2	1	0%	25%							0%	9%
1417	0	1	3	0%	17%							0%	4%	1	0	0	100%	25%							100%	9%
1421	0	0	6	0%	0%	0	0	1				0%	0%	1	0	2	33%	25%	0	0	1				25%	9%
1440	0	1	4	0%	17%							0%	4%	0	2	2	0%	25%	0	0	1				0%	9%
1441	0	2	3	0%	33%							0%	8%	0	3	2	0%	38%							0%	14%
1443	3	0	0	100%	100%	1	0	0				100%	31%	2	0	1	67%	50%	1	0	0				75%	27%
1448	0	0	9	0%	0%	0	0	1				0%	0%	0	1	1	0%	13%	0	1	0				0%	9%
1522	0	0	2	0%	0%	0	1	1				0%	4%	1	0	1	50%	25%	0	0	2				25%	9%
1527	0	1	3	0%	17%	0	1	2	0	0	2	0%	8%	0	1	5	0%	13%	0	0	2				0%	5%
Avgs.	0.3	0.6	3.6	11%	20%	0.2	0.4	1.0	0.0	0.0	2.0	11%	6%	0.6	1	1.7	28%	26%	0.2	0.2	1.0	###	###	###	25%	11%
Experts	3					5			5					4					3			4				
Treatment Group 2nd Evaluation, Category 4.0 - Human Resource Development & Management																										
	Item 4.1					Areas for			Site Visit			Item 4.1		Item 4.2					Areas for			Site Visit			Item 4.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2402	1	1	4	17%	50%							17%	12%	2	0	3	40%	50%							40%	18%
2405	0	1	5	0%	17%	0	0	2	0	0	1	0%	4%	1	0	2	33%	25%	0	0	2	0	1	0	17%	14%
2419	1	1	4	17%	50%	0	0	2				13%	12%	0	3	1	0%	38%	0	0	1	0	1	0	0%	18%
2424	0	1	2	0%	17%	0	0	2				0%	4%	0	1	1	0%	13%	1	0	0				33%	14%
2433	0	2	2	0%	33%	1	1	3	1	0	2	17%	27%	0	4	1	0%	50%	0	0	3	0	1	1	0%	23%
2439	0	2	4	0%	33%	1	0	3	1	0	1	17%	23%	0	2	3	0%	25%	1	0	1	0	0	2	11%	18%
2449	3	0	2	60%	100%	0	0	2	0	1	2	30%	27%	1	2	3	17%	50%	1	0	1	0	0	2	20%	27%
2504	1	2	4	14%	67%	0	1	0	0	1	0	11%	23%	0	1	3	0%	13%	1	0	0				20%	14%
2508	0	1	3	0%	17%	0	1	1	0	1	1	0%	12%	1	0	2	33%	25%	0	0	2	0	0	2	14%	9%
2546	1	0	4	20%	33%	0	1	1	0	0	1	13%	12%	1	0	1	50%	25%							50%	9%
Avgs.	0.7	1.1	3.4	13%	42%	0.2	0.4	1.8	0.3	0.4	1.1	12%	15%	0.6	1.3	2	17%	31%	0.5	0	1.3	0	0.5	1.2	21%	16%
Overall	0.5	0.84	3.5	12%	32%	0.2	0.4	1.5	0.3	0.4	1.3	11%	11%	0.6	1.2	1.8	22%	29%	0.4	0.1	1.1	0	0.5	1.2	23%	14%
Avgs.																										

C2, T2-C4.0, Q10CANAL.XLS

Count Data from Q10 Content Analysis for Category 4.0

Control																										
	Item 4.3					Areas for			Site Visit			Item 4.3		Item 4.4					Areas for			Site Visit			Item 4.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1413	0	1	0	0%	10%	0	0	1				0%	4%	1	0	0	100%	25%							100%	8%
1417	0	1	0	0%	10%							0%	4%	0	1	1	0%	13%							0%	4%
1421	1	1	1	33%	30%	0	0	1				25%	11%	0	3	0	0%	38%	0	0	1				0%	13%
1440	1	1	1	33%	30%	0	0	2	0	0	1	17%	11%	0	3	3	0%	38%	0	0	1				0%	13%
1441	2	0	1	67%	40%							67%	14%	0	1	1	0%	13%	0	0	3				0%	4%
1443	2	1	1	50%	50%	1	0	0				60%	25%	2	0	1	67%	50%	1	0	0	1	0	0	80%	33%
1448	2	0	3	40%	40%	0	0	1				33%	14%	1	2	2	20%	50%	0	0	4				11%	17%
1522	1	1	2	25%	30%	0	0	2				17%	11%	2	1	1	50%	63%	1	0	0				60%	29%
1527	0	0	4	0%	0%	0	0	1				0%	0%	0	0	3	0%	0%	0	0	2	1	0	0	17%	8%
Avgs.	**1.0**	**0.7**	**1.4**	**28%**	**27%**	**0.1**	**0.0**	**1.1**	**0.0**	**0.0**	**1.0**	**24%**	**10%**	**0.7**	**1.2**	**1.3**	**26%**	**32%**	**0.3**	**0.0**	**1.6**	**1.0**	**0.0**	**0.0**	**30%**	**14%**
Experts	5					3			6					4					4			4				
Treatme																										
	Item 4.3					Areas for			Site Visit			Item 4.3		Item 4.4					Areas for			Site Visit			Item 4.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2402	2	2	2	33%	60%							33%	21%	2	0	2	50%	50%	0	0	3				29%	17%
2405	1	1	0	50%	30%	0	0	1	0	0	1	25%	11%	1	0	2	33%	25%	0	1	0	0	0	1	20%	13%
2419	0	2	2	0%	20%	1	0	2	0	0	2	11%	14%	2	1	0	67%	63%	1	1	0				60%	33%
2424	1	0	2	33%	20%	0	0	1				25%	7%	1	1	1	33%	38%	0	1	1				20%	17%
2433	1	1	3	20%	30%	0	0	2	0	0	1	13%	11%	1	1	1	33%	38%	0	1	0	0	0	1	20%	17%
2439	1	1	3	20%	30%	0	0	2	0	0	1	13%	11%	0	1	2	0%	13%	0	0	1	1	0	0	20%	13%
2449	2	1	3	33%	50%	1	0	1	1	0	0	44%	32%	1	2	4	14%	50%	0	0	2	0	0	3	8%	17%
2504	1	1	3	20%	30%	0	0	1				17%	11%	1	0	2	33%	25%	0	0	1				25%	8%
2508	1	0	7	13%	20%	0	0	3	0	1	2	7%	11%	2	1	1	50%	63%	0	0	1	0	0	2	29%	21%
2546	0	3	1	0%	30%	0	0	1				0%	11%	0	2	2	0%	25%							0%	8%
Avgs.	**1.0**	**1.2**	**2.6**	**22%**	**32%**	**0.2**	**0**	**1.6**	**0.2**	**0.2**	**1.2**	**19%**	**14%**	**1.1**	**0.9**	**1.7**	**31%**	**39%**	**0.1**	**0.4**	**1**	**0.2**	**0**	**1.4**	**23%**	**16%**
Overall	**1.0**	**0.9**	**2.1**	**25%**	**29%**	**0.2**	**0**	**1.4**	**0.1**	**0.1**	**1.1**	**21%**	**12%**	**0.9**	**1.1**	**1.5**	**29%**	**36%**	**0.2**	**0.3**	**1.3**	**0.4**	**0**	**1.0**	**26%**	**15%**
Avgs.																										

Count Data from Q10 Content Analysis for Category 5.0

Control Group 2nd Evaluation, Category 5.0 - Process Management																										
	Item 5.1					Areas for			Site Visit			Item 5.1		Item 5.2					Areas for			Site Visit			Item 5.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	2	2	2	33%	75%							33%	23%	0	1	5	0%	13%							0%	4%
1415	1	1	0	50%	38%							50%	12%	1	0	1	50%	25%							50%	8%
1425	0	2	5	0%	25%	0	0	1	0	1	1	0%	12%	0	0	4	0%	0%	0	1	0				0%	4%
1436	3	1	12	19%	88%	0	0	2	0	1	2	14%	31%	2	0	11	15%	50%	0	0	2	0	0	1	13%	17%
1501	3	0	2	60%	75%	0	0	1	0	1	0	43%	27%	3	0	3	50%	75%				0	0	1	43%	25%
1506	0	0	5	0%	0%	0	0	2	0	0	1	0%	0%	0	0	5	0%	0%	0	1	0				0%	4%
1527	0	1	4	0%	13%	0	1	0	0	0	1	0%	8%	0	0	4	0%	0%	0	0	2				0%	0%
1534	0	4	3	0%	50%	0	0	1				0%	15%	0	0	2	0%	0%				1	0	0	33%	8%
1541	0	1	4	0%	13%	0	1	1				0%	8%	0	1	5	0%	13%							0%	4%
1554	0	0	1	0%	0%	0	2	0	0	0	1	0%	8%	0	1	1	0%	13%	1	1	2	1	0	2	22%	25%
Avgs.	0.9	1.2	3.8	16%	38%	0.0	0.5	1.0	0.0	0.5	1.0	14%	14%	0.6	0.3	4.1	12%	19%	0.2	0.6	1.2	0.5	0.0	1.0	16%	10%
Experts	4					4			5					4					3			5				
Treatment Group 2nd Evaluation, Category 5.0 - Process Management																										
	Item 5.1					Areas for			Site Visit			Item 5.1		Item 5.2					Areas for			Site Visit			Item 5.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2412	1	1	2	25%	38%	0	0	1				20%	12%	0	1	1	0%	13%	1	1	0	1	0	1	33%	25%
2423	1	1	3	20%	38%	1	1	1	1	1	2	25%	35%	0	0	2	0%	0%				0	0	1	0%	0%
2424	0	1	2	0%	13%	0	0	1				0%	4%	0	0	2	0%	0%	0	0	1				0%	0%
2429	1	2	2	20%	50%				0	1	2	13%	19%	0	1	1	0%	13%	0	0	1	0	0	2	0%	4%
2439	0	0	5	0%	0%	0	1	1	0	0	1	0%	4%	0	2	1	0%	25%	1	1	2	0	0	2	11%	21%
2505	1	1	5	14%	38%	0	0	2	0	0	1	10%	12%	0	0	3	0%	0%	1	0	2	0	0	1	14%	8%
2507	0	4	2	0%	50%	0	0	2				0%	15%	0	0	3	0%	0%	0	1	3				0%	4%
2509	0	3	9	0%	38%							0%	12%	0	1	5	0%	13%	0	0	1	0	0	1	0%	4%
2536	0	1	3	0%	13%	0	1	1	0	0	1	0%	8%	0	1	2	0%	13%	0	1	0				0%	8%
2537	1	2	3	17%	50%	0	0	2				13%	15%	0	1	4	0%	13%	0	1	2				0%	8%
2548	0	1	5	0%	13%	2	0	2				20%	19%	0	1	3	0%	13%	0	0	1				0%	4%
2550	0	0	5	0%	0%							0%	0%	0	0	3	0%	0%	0	1	1	0	0	1	0%	4%
2553	2	2	4	25%	75%	0	0	1				22%	23%	2	1	2	40%	63%	0	0	2				29%	21%
Avgs.	0.5	1.5	3.8	9%	32%	0.3	0.3	1.4	0.2	0.4	1.4	9%	14%	0.2	0.7	2.5	3%	13%	0.3	0.5	1.3	0.1	0.0	1.3	7%	9%
Overall	0.7	1.3	3.8	12%	34%	0.2	0.4	1.2	0.1	0.5	1.2	11%	14%	0.3	0.5	3.2	7%	15%	0.2	0.5	1.3	0.3	0.0	1.2	11%	9%
Avgs.																										

C2, T2-5.0, Q10CANAL.XLS

Count Data from Q10 Content Analysis for Category 5.0

Control																										
	Item 5.3					Areas for			Site Visit			Item 5.3		Item 5.4					Areas for			Site Visit			Item 5.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1403	1	2	6	11%	50%	0	1	1	0	0	1	8%	21%	0	2	3	0%	20%							0%	8%
1415	0	0	3	0%	0%							0%	0%	1	0	1	50%	20%							50%	8%
1425	0	1	3	0%	13%	0	1	0	0	0	1	0%	8%	1	1	0	50%	30%	0	0	1	1	0	1	40%	21%
1436	1	0	2	33%	25%							33%	8%	3	1	8	25%	70%	0	0	1				23%	29%
1501	1	0	8	11%	25%				0	0	1	10%	8%	2	2	4	25%	60%				0	1	0	22%	29%
1506	0	0	3	0%	0%	1	0	1				20%	8%	0	2	2	0%	20%	0	0	2				0%	8%
1527	0	1	2	0%	13%	0	0	2				0%	4%	0	1	1	0%	10%	0	0	2				0%	4%
1534	0	1	3	0%	13%				0	0	1	0%	4%	2	2	3	29%	60%				0	0	1	25%	25%
1541	0	0	4	0%	0%							0%	0%	1	1	1	33%	30%							33%	13%
1554	0	0	3	0%	0%	0	1	1	0	0	1	0%	4%	0	0	2	0%	0%	0	0	1	0	0	1	0%	0%
Avgs.	0.3	0.5	3.7	6%	14%	0.2	0.6	1.0	0.0	0.0	1.0	7%	7%	1	1.2	2.5	21%	32%	0.0	0.0	1.4	0.3	0.3	0.8	19%	15%
Experts	4					4			4					5					1			6				
Treatme																										
	Item 5.3					Areas for			Site Visit			Item 5.3		Item 5.4					Areas for			Site Visit			Item 5.4	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2412	0	1	1	0%	13%	0	0	2	0	0	1	0%	4%	0	1	1	0%	10%	0	0	2	0	0	1	0%	4%
2423	0	0	1	0%	0%	0	1	2	0	1	1	0%	8%	3	0	0	100%	60%				0	1	0	75%	29%
2424	0	1	1	0%	13%	1	0	1				25%	13%	0	1	1	0%	10%	0	0	2				0%	4%
2429	0	1	2	0%	13%				0	1	0	0%	8%	0	2	0	0%	20%	0	0	1	0	1	0	0%	13%
2439	0	1	2	0%	13%	1	2	0	0	0	1	14%	21%	1	2	2	20%	40%	0	0	3	0	1	0	11%	21%
2505	1	0	2	33%	25%	2	0	1	1	1	1	44%	38%	1	1	0	50%	30%	0	0	4	0	0	1	14%	13%
2507	0	1	4	0%	13%	0	2	1				0%	13%	1	2	1	25%	40%	0	0	1				20%	17%
2509	0	2	11	0%	25%	0	0	1				0%	8%	2	0	1	67%	40%	0	0	1				50%	17%
2536	0	1	5	0%	13%	0	0	1	0	1	0	0%	8%	1	1	0	50%	30%							50%	13%
2537	0	0	6	0%	0%	0	2	1				0%	8%	0	2	3	0%	20%	0	0	3				0%	8%
2548	1	0	1	50%	25%	0	1	3				17%	13%	0	3	2	0%	30%	0	0	4				0%	13%
2550	0	0	3	0%	0%	0	0	1				0%	0%	1	1	0	50%	30%	0	0	1	0	0	1	25%	13%
2553	0	2	4	0%	25%	0	1	0				0%	13%	0	4	1	0%	40%							0%	17%
Avgs.	0.2	0.8	3.3	6%	13%	0.3	0.8	1.2	0.2	0.7	0.7	8%	12%	0.8	1.5	0.9	28%	31%	0.0	0.0	2.2	0.0	0.5	0.5	19%	14%
Overall	0.2	0.7	3.5	6%	14%	0.3	0.7	1.1	0.1	0.4	0.8	7%	10%	0.9	1.4	1.6	25%	31%	0.0	0.0	1.9	0.1	0.4	0.6	19%	14%
Avgs.																										

C2, T2-5.0, Q10CANAL.XLS

Count Data from Q10 Content Analysis for Category 6.0

Control Group 2nd Evaluation, Category 6.0 - Business Results

	Item 6.1					Areas for			Site Visit			Item 6.1	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1410	0	2	1	0%	25%	0	0	1	0	0	2	0%	8%
1420	0	1	0	0%	13%	1	0	1				33%	13%
1422	1	0	2	33%	25%	0	0	4				14%	8%
1425	1	0	2	33%	25%	1	0	1				40%	17%
1430	1	0	1	50%	25%	0	0	1				33%	8%
1436	0	1	0	0%	13%	0	0	1	0	0	1	0%	4%
1441	0	2	1	0%	25%	0	0	3				0%	8%
1442	1	1	3	20%	38%							20%	13%
1501	0	0	1	0%	0%	1	0	0	1	0	0	67%	17%
1513	1	0	2	33%	25%	0	0	2	0	0	2	14%	8%
1543	2	1	0	67%	63%	0	0	2	0	0	2	29%	21%
1552	2	0	0	100%	50%	1	0	0	0	1	0	75%	29%
Avgs.	**0.8**	**0.7**	**1.1**	**28%**	**27%**	**0.4**	**0.0**	**1.5**	**0.2**	**0.2**	**1.2**	**27%**	**13%**
Experts	4					4			4				

	Item 6.2					Areas for			Site Visit			Item 6.2	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1410	2	1	2	40%	83%	0	0	1	0	0	1	29%	23%
1420	0	1	0	0%	17%	1	0	1				33%	14%
1422	0	1	1	0%	17%	0	0	3				0%	5%
1425	0	1	1	0%	17%	0	0	1				0%	5%
1430	0	0	4	0%	0%	1	0	0				20%	9%
1436	1	0	0	100%	33%	1	0	3	1	0	1	43%	27%
1441	2	1	0	67%	83%	1	0	1				60%	32%
1442	1	2	0	33%	67%	0	0	4				14%	18%
1501	0	0	1	0%	0%	1	0	0	1	0	1	50%	18%
1513	0	1	1	0%	17%	0	0	2	0	0	2	0%	5%
1543	2	0	0	100%	67%	0	0	3	0	0	1	33%	18%
1552	0	1	1	0%	17%	1	0	1	1	0	0	40%	23%
Avgs.	**0.7**	**0.8**	**0.9**	**28%**	**35%**	**0.5**	**0.0**	**1.7**	**0.5**	**0.0**	**1.0**	**27%**	**16%**
Experts	3					4			4				

	Item 6.3					Areas for			Site Visit			Item 6.3	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1410	0	1	1	0%	10%	0	0	2	0	0	1	0%	4%
1420	0	1	0	0%	10%	1	0	0				50%	13%
1422	2	0	1	67%	40%	0	0	1				50%	17%
1425	1	0	2	33%	20%	0	0	3				17%	8%
1430	0	1	2	0%	10%	0	0	1				0%	4%
1436	4	0	1	80%	80%	1	0	0				83%	42%
1441	2	1	1	50%	50%	0	0	1				40%	21%
1442	2	1	3	33%	50%	0	0	1				29%	21%
1501	0	0	3	0%	0%	1	0	0	0	0	2	17%	8%
1513	0	1	3	0%	10%	0	0	1	0	0	1	0%	4%
1543	3	0	1	75%	60%							75%	25%
1552	1	0	1	50%	20%	1	0	0	0	0	1	50%	17%
Avgs.	**1.3**	**0.5**	**1.6**	**32%**	**30%**	**0.4**	**0.0**	**0.9**	**0.0**	**0.0**	**1.3**	**34%**	**15%**
Experts	5					3			4				

Treatment Group 2nd Evaluation, Category 6.0 - Business Results

	Item 6.1					Areas for			Site Visit			Item 6.1	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2407	1	0	0	100%	25%	1	0	0	0	0	1	67%	17%
2411	1	1	1	33%	38%	1	0	2	1	0	1	38%	29%
2414	1	0	3	25%	25%	1	0	5	0	0	2	17%	17%
2429	1	0	0	100%	25%	1	0	0				100%	17%
2433	0	1	1	0%	13%	1	0	2	0	1	0	17%	17%
2434	1	0	1	50%	25%	1	0	1	1	0	1	50%	25%
2435	2	1	0	67%	63%	1	0	0	1	0	0	80%	38%
2548	0	1	0	0%	13%	1	0	2				25%	13%
2551	0	0	1	0%	0%	1	0	3				20%	8%
Avgs.	**0.8**	**0.4**	**0.8**	**42%**	**25%**	**1.0**	**0.0**	**1.7**	**0.5**	**0.2**	**0.8**	**46%**	**20%**
Overall Avgs.	**0.8**	**0.6**	**1**	**34%**	**26%**	**0.7**	**0**	**1.6**	**0.3**	**0.2**	**1**	**35%**	**16%**

	Item 6.2					Areas for			Site Visit			Item 6.2	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2407	1	0	0	100%	33%	1	0	0	0	0	1	67%	18%
2411	2	0	0	100%	67%	0	0	2	0	0	3	29%	18%
2414	0	1	5	0%	17%	1	0	4	1	1	0	15%	27%
2429	0	1	1	0%	17%	0	0	1				0%	5%
2433	0	1	1	0%	17%	0	0	3				0%	5%
2434	0	1	1	0%	17%	1	0	2	1	0	1	29%	23%
2435	2	1	0	67%	83%	1	0	1	0	1	1	43%	36%
2548	0	1	0	0%	17%	1	0	2	0	0	2	17%	14%
2551	0	1	1	0%	17%	1	0	3				17%	14%
Avgs.	**0.6**	**0.8**	**1.0**	**30%**	**31%**	**0.7**	**0.0**	**2.0**	**0.3**	**0.3**	**1.3**	**24%**	**18%**
Overall Avgs.	**0.6**	**0.8**	**1**	**29%**	**33%**	**0.6**	**0**	**1.8**	**0.4**	**0.2**	**1.2**	**26%**	**17%**

	Item 6.3					Areas for			Site Visit			Item 6.3	
	Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2407	0	0	1	0%	0%	1	0	0				50%	8%
2411	0	3	0	0%	30%	0	0	1				0%	13%
2414	1	0	2	33%	20%	1	0	1	0	0	2	29%	17%
2429	1	0	0	100%	20%	1	0	0				100%	17%
2433	2	0	0	100%	40%	1	0	0				100%	25%
2434	2	0	1	67%	40%	0	1	1	0	1	1	29%	25%
2435	0	2	1	0%	20%	1	1	0				20%	21%
2548	0	1	1	0%	10%	1	1	1				20%	17%
2551					0%	1	0	4				20%	8%
Avgs.	**0.8**	**0.8**	**0.8**	**38%**	**20%**	**0.8**	**0.3**	**0.9**	**0.0**	**0.5**	**1.5**	**41%**	**17%**
Overall Avgs.	**1.1**	**0.6**	**1.3**	**34%**	**26%**	**0.6**	**0.2**	**0.9**	**0.0**	**0.2**	**1.3**	**37%**	**16%**

Count Data from Q10 Content Analysis for Category 7.0

Control Group 2nd Evaluation, Category 7.0 - Customer Focus & Satisfaction																										
	Item 7.1					Areas for			Site Visit			Item 7.1		Item 7.2					Areas for			Site Visit			Item 7.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
1413	1	1	2	25%	30%							25%	13%	2	1	1	50%	50%	0	0	1				40%	19%
1431	2	0	3	40%	40%	0	0	1				33%	17%	1	2	1	25%	40%							25%	15%
1448	0	2	3	0%	20%	0	1	0	0	0	1	0%	13%	0	1	5	0%	10%							0%	4%
1513	1	2	1	25%	40%	1	0	0	0	0	1	33%	25%	2	0	2	50%	40%	0	1	1	0	0	1	29%	19%
1525	0	1	1	0%	10%	0	0	4	0	0	2	0%	4%	1	1	2	25%	30%	0	0	2	0	0	1	14%	12%
1530	2	1	3	33%	50%	0	0	2				25%	21%	0	3	2	0%	30%	0	0	2				0%	12%
1541	1	0	2	33%	20%							33%	8%	0	1	4	0%	10%							0%	4%
1544	1	1	0	50%	30%	0	0	1				33%	13%	0	1	2	0%	10%							0%	4%
1555	3	2	5	30%	80%	0	0	1	0	0	1	25%	33%	1	2	3	17%	40%	0	0	1	0	0	1	13%	15%
Avgs.	**1.2**	**1.1**	**2.2**	**26%**	**36%**	**0.1**	**0.1**	**1.3**	**0.0**	**0.0**	**1.3**	**23%**	**16%**	**0.8**	**1.3**	**2.4**	**19%**	**29%**	**0.0**	**0.2**	**1.4**	**0.0**	**0.0**	**1.0**	**13%**	**12%**
Experts	5					3			4					5					3			5				

Treatment Group 2nd Evaluation, Category 7.0 - Customer Focus & Satisfaction																										
	Item 7.1					Areas for			Site Visit			Item 7.1		Item 7.2					Areas for			Site Visit			Item 7.2	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp
2408	2	2	1	40%	60%	0	0	1	0	0	1	29%	25%	1	2	2	20%	40%	0	0	1	0	0	1	14%	15%
2414	1	1	4	17%	30%	0	0	2	0	0	2	10%	13%	1	2	4	14%	40%	0	0	2	0	0	2	9%	15%
2423	1	1	1	33%	30%	0	0	7	1	0	0	18%	21%	1	3	2	17%	50%	0	1	0	2	0	0	33%	38%
2427	1	0	2	33%	20%							33%	8%	0	1	2	0%	10%				0	0	1	0%	4%
2434	1	0	2	33%	20%	0	0	2	0	0	1	17%	8%	1	0	2	33%	20%	1	0	1	0	0	2	29%	15%
2504	2	2	3	29%	60%	0	0	1				25%	25%	3	1	2	50%	70%							50%	27%
2535	1	1	1	33%	30%							33%	13%	1	1	2	25%	30%							25%	12%
2546	1	0	7	13%	20%							13%	8%	0	4	2	0%	40%							0%	15%
2551	1	1	4	17%	30%	0	0	2	0	0	2	10%	13%	1	0	5	17%	20%							17%	8%
Avgs.	**1.2**	**0.9**	**2.8**	**28%**	**33%**	**0.0**	**0.0**	**2.5**	**0.2**	**0.0**	**1.2**	**21%**	**15%**	**1.0**	**1.6**	**2.6**	**20%**	**36%**	**0.3**	**0.3**	**1.0**	**0.4**	**0.0**	**1.2**	**20%**	**17%**
Overall	**1.2**	**1.0**	**2.5**	**27%**	**34%**	**0.1**	**0.1**	**1.8**	**0.1**	**0.0**	**1.2**	**22%**	**16%**	**0.9**	**1.4**	**2.5**	**19%**	**32%**	**0.1**	**0.2**	**1.2**	**0.3**	**0.0**	**1.1**	**17%**	**14%**
Avgs.																										

Count Data from Q10 Content Analysis for Category 7.0

Control																																							
	Item 7.3					Areas for			Site Visit			Item 7.3		Item 7.4					Areas for			Site Visit			Item 7.4		Item 7.5					Areas for			Site Visit			Item 7.5	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Ex
1413	0	1	0	0%	10%							0%	3%					0%	0	0	1				0%	0%	1	0	0	100%	33%							100%	8%
1431	1	0	4	20%	20%							20%	7%	0	2	0	0%	20%	0	0	1				0%	9%	0	0	2	0%	0%							0%	0%
1448	0	1	3	0%	10%	0	0	1				0%	3%	2	1	1	50%	50%							50%	23%	1	0	1	50%	33%	1	0	0				67%	17%
1513	1	1	1	33%	30%	0	1	0	0	0	1	20%	13%	2	0	1	67%	40%	0	0	2				40%	18%	1	0	1	50%	33%	0	0	1				33%	8%
1525	1	1	2	25%	30%	0	0	1	0	0	2	14%	10%	1	1	0	50%	30%	1	0	2	0	1	0	33%	27%	1	0	0	100%	33%	1	0	1	1	0	0	75%	25%
1530	3	0	3	50%	60%	0	0	1				43%	20%	2	0	2	50%	40%	0	0	1				40%	18%	1	0	2	33%	33%	0	0	1				25%	8%
1541	0	0	1	0%	0%	0	0	1				0%	0%	0	1	2	0%	10%							0%	5%	0	0	1	0%	0%							0%	0%
1544	0	2	0	0%	20%	0	0	1				0%	7%	1	0	1	50%	20%	0	1	0				33%	14%	1	0	0	100%	33%	0	0	1				50%	8%
1555	1	1	3	20%	30%	0	1	2				13%	13%	0	1	1	0%	10%	0	0	1	0	0	1	0%	5%	1	0	2	33%	33%	0	0	1				25%	8%
Avgs.	**0.8**	**0.8**	**1.9**	**16%**	**23%**	**0.0**	**0.3**	**1.0**	**0.0**	**0.0**	**1.5**	**12%**	**9%**	**1.0**	**0.8**	**1.0**	**33%**	**24%**	**0.1**	**0.1**	**1.1**	**0.0**	**0.5**	**0.5**	**22%**	**13%**	**0.8**	**0.0**	**1.0**	**52%**	**26%**	**0.3**	**0.0**	**0.8**	**1.0**	**0.0**	**0.0**	**42%**	**9%**
Experts	5					6			4					5					4			2					3					5			4				
Treatme																																							
	Item 7.3					Areas for			Site Visit			Item 7.3		Item 7.4					Areas for			Site Visit			Item 7.4		Item 7.5					Areas for			Site Visit			Item 7.5	
	Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages		Strengths					Improvement			Issues			Averages	
Subject	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Exp	H	A	M	%H	%Exp	H	A	M	H	A	M	%H	%Ex
2408	2	1	1	50%	50%	0	0	1	0	0	2	29%	17%	0	1	1	0%	10%	0	0	1	0	0	1	0%	5%	0	0	1	0%	0%				0	0	1	0%	0%
2414	1	1	4	17%	30%	0	1	2	0	1	1	9%	17%	2	1	1	50%	50%	0	0	2	0	0	1	29%	23%	1	0	1	50%	33%	0	0	2				25%	8%
2423	0	1	0	0%	10%	1	0	1	1	0	1	40%	17%	0	1	1	0%	10%	0	1	2	0	0	1	0%	9%	1	0	0	100%	33%	0	1	2	0	1	2	14%	17%
2427	0	1	1	0%	10%	0	1	0	0	0	1	0%	7%					0%	0	0	2	0	0	1	0%	0%	1	0	0	100%	33%	0	1	0				50%	13%
2434	0	1	1	0%	10%	2	0	1	1	0	1	43%	23%	2	0	0	100%	40%	0	1	2	0	0	3	25%	23%	1	1	0	50%	50%	0	1	0	0	0	2	20%	17%
2504	2	1	1	50%	50%	1	0	1				50%	23%	2	1	0	67%	50%	0	0	1				50%	23%	1	0	1	50%	33%	0	0	1				33%	8%
2535	0	2	1	0%	20%	0	0	1				0%	7%	2	1	0	67%	50%							67%	23%	1	0	1	50%	33%	0	0	1	0	0	1	25%	8%
2546	0	1	3	0%	10%	0	0	1				0%	3%	1	0	1	50%	20%	0	0	1				33%	9%	0	0	3	0%	0%							0%	0%
2551	0	2	1	0%	20%	0	1	2	0	0	1	0%	10%	0	0	1	0%	0%	0	0	1				0%	0%	0	0	1	0%	0%	1	0	2				25%	8%
Avgs.	**0.6**	**1.2**	**1.4**	**13%**	**23%**	**0.4**	**0.3**	**1.1**	**0.3**	**0.2**	**1.2**	**19%**	**14%**	**1.1**	**0.6**	**0.6**	**42%**	**26%**	**0.0**	**0.3**	**1.5**	**0.0**	**0.0**	**1.4**	**23%**	**13%**	**0.7**	**0.1**	**0.9**	**44%**	**24%**	**0.1**	**0.4**	**1.1**	**0.0**	**0.3**	**1.5**	**21%**	**9%**
Overall	**0.7**	**1.0**	**1.7**	**15%**	**23%**	**0.3**	**0.3**	**1.1**	**0.3**	**0.1**	**1.3**	**16%**	**11%**	**1.1**	**0.7**	**0.8**	**38%**	**25%**	**0.1**	**0.2**	**1.3**	**0**	**0.1**	**1.1**	**22%**	**13%**	**0.7**	**0.1**	**0.9**	**48%**	**25%**	**0.2**	**0.2**	**1.0**	**0.2**	**0.2**	**1.2**	**32%**	**9%**
Avgs.																																							

Appendix AV. Test Statistic and Rejection Region for H4.

Test Statistic

$$F_{\text{max obs.}} = \frac{s^2_{\text{max}}}{s^2_{\text{min}}}$$

where s^2_{max} is the square of the standard deviation of the scores on the category with the largest standard deviation and s^2_{min} is the square of the standard deviation of the scores on the category with the smallest standard deviation. Standard deviations were taken from the Edited Minitab Session Files for Hypothesis 4 (see Appendix B).

Rejection Region

The null hypothesis, H_0, was rejected if $F_{\text{max obs.}}$ was greater than $F_{\text{max critical}}$.

$$F_{\text{max critical}} = F_{\text{max}(t,\ df_2)0.95} = F_{\text{max}(7,\ 23)0.95}$$

where t = number of populations (categories) being compared, df_2 = average[1] category sample size minus one or $n_{\text{average}} - 1$. When actual df were not available in the table, the next lowest df in the table were used. Below are the two F_{max} values closest to the actual degrees of freedom of $F_{\text{max critical}}$.

$F_{\text{max}(7,\ 20)0.95} = 3.94$ and $F_{\text{max}\ (7,\ 30)0.95} = 3.02$

(from Ott, 1984, p. 723; Pearson and Hartley, 1970, p. 202)

[1] When dealing with unequal sample sizes, Ott (1984) recommends the use of the maximum sample size minus one. Schulman (1996) considers this too "anti-conservative" (i.e., increases probability of a Type I error too much) and recommended the use of the average sample size minus one. Pearson and Hartley (1970, p. 67) also recommend the use of the average sample size minus one.

Appendix AW. Test Statistic and Rejection Region for H9.

Test Statistic

$$F_{\text{max obs.}} = \frac{s^2{}_{\text{max}}}{s^2{}_{\text{min}}}$$

where $s^2{}_{\text{max}}$ is the square of the standard deviation of the scores on the category with the largest standard deviation and $s^2{}_{\text{min}}$ is the square of the standard deviation of the scores on the category with the smallest standard deviation. Standard deviations were taken from Appendix F: Edited Minitab Session Files of Descriptive Statistics for Hypothesis 6.

Rejection Region

The null hypothesis, H_0, was rejected if $F_{\text{max obs.}}$ was greater than $F_{\text{max critical}}$.

$F_{\text{max critical}} = F_{\text{max}(t,\, df_2)0.95} = F_{\text{max}(7,\, 9)0.95} = 8.41$ (Ott, 1984, p. 723)

where t= number of populations (categories) being compared,

df_2 = average[2] category sample size minus one or $n_{\text{average}} - 1$

The df were coincidentally the same for both the treatment and control group.

[2] When dealing with unequal sample sizes, Ott (1984) recommends the use of the maximum sample size minus one. Schulman (1996) considers this too "anti-conservative" (i.e., increases probability of a Type I error too much) and recommended the use of the average sample size minus one. Pearson and Hartley (1970, p. 67) also recommend the use of the average sample size minus one.

Appendix AX. Test Statistics and Rejection Regions for H10.

Test Statistics

Minitab was used to calculate the F-statistics for the two-factor ANOVA using the general linear model (GLM). The GLM was used due to the unbalanced design. Minitab was used to calculate the t-statistic using a pooled variance estimate, s_p^2. Under the null hypothesis, it is reasonable to expect the variance of the accuracy indices to be the same. Minitab was also used to the calculate the W-statistic for the Mann-Whitney tests. The Friedman-type rank test required ranking observations within rows for testing column effects and vice versa. The Friedman-type rank test also required the calculation of a vector of modified sums of ranks, R, and the null covariance matrix of R, Σ. These rankings and calculations were done by constructing an Excel spreadsheet for each category, accuracy index, and main effect (see Appendix K for an example spreadsheet). The test statistic for the Friedman-type rank test, T, was calculated using the following formula (Mack and Skillings, 1980, p. 947).

$$T = R'\Sigma^{-}R$$

"The T statistic has a limiting (N → ∞) chi-squared distribution having J-1 degrees of freedom" (Mack and Skillings, 1980, p. 948). The generalized inverse of the null covariance matrix of R was calculated using SAS and then inserted into the Excel spreadsheet. The matrix algebra for calculating T was performed in the Excel spreadsheet (see bottom of spreadsheet in Appendix K).

Rejection Regions

Rather than simply test the observed value against a tabularized critical value, Minitab was used to calculate the actual p-value for each test statistic (F_{obs}, t_{obs}, T_{obs}, or W_{obs}). For F-tests, the p-value represents the probability that F would be at least as large as F_{obs} when the null hypothesis is true. For the one-sided t-tests, the p-value represents the probability that the value of t would be at least as large as the value of t_{obs} when the null hypothesis is true. For the Friedman-type rank tests, the p-value represents the probability that the value of T would be at least as large as the value of T_{obs} when the null hypothesis is true. For the one-sided Mann-Whitney tests, the p-value represents the probability that the value of W would be at least as large as the value of W_{obs} when the null hypothesis is true.

Appendix AY. Observations from the Correlation Analysis

The following comments were prepared after reviewing the correlation table. Strong correlations that were expected are not discussed unless there are implications for further analysis.

- Elevation and dimensional accuracy are negatively correlated (r = -0.435). Those who had better dimensional accuracy tended to be more lenient in scoring (resulting in poor elevation accuracy). I suspect those with more experience and maturity gave higher scores, yet were able to give more specific feedback regarding relative strengths and weaknesses.
- Dimensional accuracy was positively correlated with the operations job function (oper) (r = 0.364). That is, those in production or service delivery, maintenance, or QC/QA functions had poorer dimensional accuracy. The reason is unclear.
- Years of work experience (exp) and age were very strongly correlated (r = 0.958). Age should be dropped from the final stepwise regression analysis since work experience is the more relevant of the two. Work experience may also have a collinearity problem with supervisory responsibility (supv) (r = 0.455).
- Years of work experience was positively correlated with organizational size (r = 0.380). Organizational size was most strongly correlated with working for the federal government (r = 0.659), is there a possible connection? (possible sample issue)
- Years of work experience was negatively correlated (r = -0.302) with the last degree completed being in engineering (engrdeg). This could be a function of older workers without engineering degrees returning to school for an engineering degree. It could also be the result of more experienced engineers having obtained their second (most recent) degrees in non-engineering disciplines (e.g., business).
- Whether the last degree completed was in engineering (engrdeg) was negatively correlated with both supervisor responsibility (r = - 0.269) and being in an executive/administrative function (r = -0.264). It is possible that those whose careers have migrated toward management are more likely to have pursued a degree other than engineering. Both supervisor responsibility and being in a executive/administrative position are positively correlated with years of work experience (r = 0.455, r = 0.476, respectively), which may explain some of the negative correlation between years of work experience and the last degree completed being in engineering.
- Years of work experience was positively correlated (r = 0.476) with being in an executive/administrative function (exec) and negatively correlated (r = -0.409) with being a full time student (std). Neither of these is surprising, but they could cause a multi-collinearity problem.
- Years of work experience was negatively correlated (r = -0.339) with working for a state or local government. This is likely due to the large number of students working for the university who described their employers as state or local government. Being a full time student and working for a state or local government had a positive correlation of r = 0.536.

Appendix AZ. Observations from the Stepwise Regression Procedures

The following bullets summarize observations based on the exploratory stepwise regression procedure described in Chapter 4.

dep. var. = elevation, replacing only one original variable at a time, sans obs. 26

- engrdeg is frequently the first variable entered into the model. When the latest degree completed was in engineering (i.e., engrdeg = 1), elevation accuracy improves (decreases) by several points.
- For those with greater than zero but less than six and a half years experience in the quality control/quality assurance/quality improvement function, nearly half(?) the variance in their elevation accuracy is predicted by whether or not they are currently working in operations (i.e., production or service delivery, maintenance, or QC/QA). Those currently working in operations were much more accurate than those who were not.
- supv appears to have some impact on prediction. In the default model, increased supervisory responsibility predicts improved (i.e., lower index) elevation accuracy. This relation is less visible once alternate variables begin suppressing observations.
- age appears to share some predictive power with other variables, preventing age from entering the equation. Once age is deleted, exp and supv become more prominent and enter the equation sooner. The overall increase in prediction is very minor (R-squared of 25.30 versus 25.01). Dropping age reinstates an observation that failed to report age (obs. 47).
- assess and gndr seem to add little or nothing to prediction
- Among the current job function descriptors, only oper (operations) appears to have any effect. Using oper3E slightly increases this effect. Perhaps the whole set should be kept due to the indicator format.
- The employer descriptors appear to have no descriptive value. Perhaps the set should be dropped.

dep. var. = elevation, replacing and reinstating variables to increase prediction
(sans obs. 26)

- Replacing qptrng with qptrng5E increases R-squared to 27.08 using four predictors. Since qptrng5E drops the extreme outlier and all those without quality/productivity improvement training in the past ten years, this may be representative of evaluators in training. That is, those training to be evaluators for a quality/productivity related assessment.
- Replacing exp with expE1 while retaining qptrng5E produces a more refined equation for evaluators in training. Not only does expE1 drop the four extreme obs. w/25 years or more experience, but it also causes the suppression of the three most extreme

values for supervisory responsibility (i.e., 35, 70, and 220 employees). This results in a slight increase in R-squared (to 27.16), but requires only three predictors. With a total of 42 observations, this model is worthy of further consideration.

- Among the few subjects with experience in QC/QA/QI, dropping the two extreme values while retaining qptrng5E and expE1 yields an equation with an R-squared of 67.11. Unfortunately, the equation has a negative intercept and each predictor is positive (i.e., predicts worse accuracy as that variable increases). This makes interpretation rather difficult[3] Only 13 observations were used for this equation, since the alternate variables suppress four observations with between 0.25 and 6.3 years in QC/QA/QI. Compare this with the equation where all 17 observations were used (R-squared = 49.01 with only one predictor).
- Among subjects from larger organizations (size4E), an R-squared of 36.97 was produced using age, engrdeg, and mfg as predictors (qptrng5E and expE1 were available for entering the model). Dropping age produces an equation with expE1 and degree as the predictors and an R-squared of 30.92 (n = 33). This implies that subjects from large organizations with more experience and education are likely to be more lenient evaluators. Reinstating exp to expand the range of experience increased both n (to 36) and R-squared (to 35.61), and produced the following equation: Eavg = 29.81 + 12.3mfg -11.1engrdeg -0.097supv. The effect of an engineering degree and supervisory responsibility are consistent with earlier results. The large degradation in elevation accuracy predicted for those working in manufacturing was likely brought out by the exclusion of subjects from small organizations. Perhaps the small number of subjects from manufacturing was of limited value until placed in the context of a smaller sample size. Since the case being evaluated was a manufacturer, the subjects from manufacturing employers may have been more appreciative of manufacturing excellence and thus more lenient[4].

[3] Taken literally, the equation predicts a subject with less than a bachelors degree, no work experience, and no assessment experience would have an elevation accuracy of approximately -6.2. Since data for the only subject with less than a bachelors degree is suppressed due to missing responses, this hypothetical situation falls outside the predictive range of this equation. The lowest possible value for degree is two (i.e., completed a bachelors), yielding a predicted elevation accuracy of approximately 8.2 if the values of the other predictors are zero.

[4] Based on the regression data alone, an argument could be made subjects from manufacturing employers were more severe and that caused their elevation accuracy to be worse than the other subjects. Examination of the scoring data shows poor elevation accuracy to correlate with higher scores (i.e., a leniency effect).

dep. var. = dimensional accuracy, replacing only one original variable at a time, sans obs. 29

- exp appears to be an important predictor. When replaced with exp3, prediction improves and fewer predictors are required. exp3 drops the four subjects with 25 or more years of experience.
- supv is an important predictor in the default model. supv7D drops the middle managers and executives and drops those without supervisory responsibility, leaving only fourteen observations. When supv7D was used in the stepwise regression, the omissions interacted with the omissions of other variables and resulted in too little data for Minitab. To compensate, all the variables including supv7D were entered into a MLR model for analysis. Variables were dropped one at a time (based on prior knowledge and sequential sums of squares). This resulted in an equation with nine predictors that explains 87.9% (R-squared adjusted) of the variance in DA for these fourteen supervisors. These nine variables were then analyzed using stepwise regression. Two predictors (supv7D and size) explained 59.3% of the variation in DA for these supervisors. Increased supervisory responsibility and larger employing organization both related to increased dimensional accuracy. Further exploration yielded a model with eight predictors that explained 90.3% of the variation in DA for these supervisors. The utility of this model is debatable due to the interactions of the variables.
- size3D drops those reporting size as zero (i.e., the unemployed) and may be more representative of evaluators in training than using size. Using size3D only reduces the sample size to 58 and yields an R-squared of 44% with eight predictors. Further analysis is needed to see if the number of predictors can be reduced.

dep. var. = dimensional accuracy, replacing and reinstating variables to increase prediction (sans obs. 29)

- Retaining exp3 as the measure of experience appears to maintain higher prediction and produce a regression equation with rational predictors.
- Retaining both exp3 and size3D reduces the sample size to 54 and yields an R-squared of 47.9% with seven predictors. As described above, this is probably a more representative sample than the default.
- The last variable entered into the default equation is stloc, which describes the subject's employer as state or local government. Subjects describing their employer as state or local government are predicted to be 2.7 points less accurate than the others. In the model with exp3 and size3D, fed enters the equation in a similar fashion. When stloc was dropped, not only did it fall from the equation but so did fed. In their place, svc and mfg entered the equation and both predicted subjects who described their employers as service or manufacturing to be more accurate than the others. With size3D omitting the unemployed, all remaining subjects must work for an employer

described as svc, mfg, fed, or stloc. This implies that those working for the private sector were more accurate (in terms of DA) than those working in the public sector.

- When stloc was dropped, the portion of variation predicted with six variables increased from 43% to 46%. This six variable equation may prove to be a better predictor (in terms of R-squared adjusted) than the seven variable equation described above.

Appendix BA. Control Group-Screened First Evaluation Scores

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
1506	1	2	40	50	30	40.0	40	60	70	56.7																							
1555	1	2	80	60	40	60.0	100	40	90	76.7																							
1413	1	2	100	80	80	86.7	90	90	100	93.3																							
1446	1	2	50	80	30	53.3	60	70	60	63.3																							
1448	1	2	90	60	90	80.0	80	70	70	73.3																							
1409	1	2	90	100	90	93.3	90	100	90	93.3																							
1436	1	2	90	100	100	96.7	90	100	90	93.3																							
1425	1	3	90	90	78	86.0					83	80	81.5																				
1513	1	5	80	70	80	76.7													70	70	80	80	75.0										
1528	1	6	90	90	80	86.7																		90	90	80	86.7						
1530	1	6	70	60	70	66.7																		60	70	80	70.0						
1502	1	6	100	92	100	97.3																		87	100	100	95.7						
1541	2	3					90	100	80	90.0	100	100	100.0																				
1442	2	4					90	90	100	93.3				100	100	100	90	97.5															
1552	2	5					50	70	50	56.7									75	85	80	50	72.5										
1431	2	6					90	90	100	93.3														90	60	90	80.0						
1417	2	7					90	60	100	83.3																		100	100	70	70	70	82.0
1501	3	4									100	100	100.0	60	100	90	100	87.5															
1554	3	4									50	80	65.0	50	40	60	20	42.5															
1415	3	4									70	80	75.0	80	70	80	80	77.5															
1410	3	4									80	90	85.0	80	100	80	100	90.0															
1525	3	5									60	20	40.0						20	20	60	50	37.5										
1443	3	5									70	60	65.0						90	80	80	80	82.5										
1441	3	5									80	60	70.0						100	90	100	90	95.0										
1527	3	7									90	90	90.0															100	100	90	80	80	90.0
1522	3	7									95	100	97.5															80	100	90	100	90	92.0
1544	4	5												80	90	80	80	82.5	60	70	70	80	70.0										
1510	4	6												80	60	90	80	77.5						90	90	90	90.0						
1543	4	7												50	90	80	80	75.0										70	80	90	80	100	84.0
1447	4	7												70	50	90	100	77.5										100	100	70	70	80	84.0
1403	4	7												50	60	80	90	70.0										90	80	90	50	90	80.0
1422	4	7												90	90	90	80	87.5										90	80	80	90	100	88.0
1420	5	7																	75	40	35	90	60.0					40	80	25	95	70	62.0
1440	5	7																	90	70	80	90	82.5					90	80	60	80	90	80.0
1430	5	7																	80	35	80	65	65.0					90	90	70	90	90	86.0
1534	6	7																						80	90	50	73.3	90	80	80	70	90	82.0
1428	6	7																						90	90	90	90.0	90	90	90	90	80	88.0
Means			80.8	77.7	72.3	76.9	80.0	78.3	83.3	80.6	79.8	78.2	79.0	71.8	77.3	83.6	81.8	78.6	73.3	62.2	73.9	75.0	71.1	83.9	84.3	82.9	83.7	85.8	88.3	75.4	80.4	85.8	83.2
Std. Dev.			18.81	17.26	25.22	18.26	19.07	19.46	17.23	14.83	16.23	24.01	18.43	17.22	21.95	10.27	22.28	14.33	23.32	24.51	17.99	16.20	16.35	11.14	13.97	16.04	9.49	16.76	9.37	19.00	13.89	9.96	7.70
EXPERTS			70	55	50	58.3	65	60	50	58.3	45	35	40.0	65	55	70	65	63.8	55	50	50	70	56.3	55	50	60	55.0	65	65	50	65	60	61.0

Appendix BB. Treatment Group-Screened First Evaluation Scores

FirstName	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
2508	1	2	80	70	80	76.7	80	90	90	86.7																							
2424	1	2	70	80	70	73.3	80	90	80	83.3																							
2449	1	2	86	79	84	83.0	71	88	91	83.3																							
2411	1	2	90	90	100	93.3	100	100	100	100.0																							
2537	1	3	80	80	70	76.7					60	70	65.0																				
2505	1	4	90	90	90	90.0								80	90	80	90	85.0															
2531	1	5	90	100	90	93.3													100	100	100	90	97.5										
2545	1	5	95	85	90	90.0													95	90	90	85	90.0										
2546	1	5	90	80	70	80.0													80	80	80	80	80.0										
2536	1	6	70	85	100	85.0																		90	85	95	90.0						
2401	1	6	80	90	70	80.0																		70	80	50	66.7						
2548	1	7	30	30	50	36.7																						70	80	50	50	70	64.0
2515	1	7	80	90	90	86.7																						80	90	90	90	90	88.0
2437	2	3					90	20	30	46.7	90	100	95.0																				
2414	2	4					90	80	80	83.3				70	90	90	70	80.0															
2504	2	5					70	60	30	53.3									50	60	50	70	57.5										
2419	2	5					60	90	70	73.3									60	80	80	70	72.5										
2433	2	5					70	60	90	73.3									80	70	60	90	75.0										
2507	2	6					70	30	50	50.0														90	70	60	73.3						
2405	2	6					80	60	70	70.0														80	50	80	70.0						
2551	3	4									70	50	60.0	80	90	90	80	85.0															
2416	3	4									60	40	50.0	50	40	50	50	47.5															
2423	3	4									90	50	70.0	70	70	80	90	77.5															
2427	3	4									70	80	75.0	90	90	80	80	85.0															
2429	3	4									90	90	90.0	100	100	90	100	97.5															
2402	3	6									90	90	90.0											80	90	80	83.3						
2439	3	6									90	90	90.0											90	100	90	93.3						
2532	3	7									40	60	50.0															50	70	50	60	50	56.0
2521	3	7									40	60	50.0															90	90	80	70	90	84.0
2553	3	7									90	80	85.0															80	80	70	60	60	70.0
2407	3	7									80	80	80.0															90	100	80	80	90	88.0
2434	4	5												60	70	70	60	65.0	50	30	60	70	52.5										
2408	4	5												80	80	70	90	80.0	90	80	80	80	82.5										
2519	4	6												80	80	100	80	85.0						90	80	90	86.7						
2535	4	6												95	90	100	95	95.0						100	80	95	91.7						
2426	4	6												80	50	70	80	70.0						60	90	70	73.3						
2550	4	7												50	50	60	70	57.5										80	70	60	70	80	72.0
2435	4	7												20	50	30	50	37.5										90	70	50	60	80	70.0
2404	5	6																	90	90	100	90	92.5	90	80	90	86.7						
2533	5	7																	70	75	80	90	78.8					80	95	80	85	75	83.0
2444	5	7																	80	70	90	80	80.0					90	90	70	80	70	80.0
2509	6	7																						90	85	89	88.0	93	93	86	95	83	90.0
2412	6	7																						80	70	80	76.7	60	90	50	60	90	70.0
Means			79.3	80.7	81.1	80.4	78.3	69.8	71.0	73.0	73.8	72.3	73.1	71.8	74.3	75.7	77.5	74.8	76.8	75.0	79.1	81.4	78.1	84.2	80.0	80.8	81.6	79.4	84.8	68.0	71.7	77.3	76.3
Std. Dev.			16.72	16.94	14.44	14.65	11.61	26.31	24.35	16.90	18.95	18.78	16.78	21.09	19.50	19.50	15.78	17.44	17.36	18.57	16.40	8.39	13.69	10.84	12.61	14.22	9.16	13.37	10.50	15.40	14.20	12.96	10.79
EXPERTS			70	55	50	58.3	65	60	50	58.3	45	35	40.0	65	55	70	65	63.8	55	50	50	70	56.3	55	50	60	55.0	65	65	50	65	60	61.0

Appendix BC. Control Group-Screened Second Evaluation Scores

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4.1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
1428	1	3	90	90	90	90.0					90	90	90.0																				
1421	1	4	80	70	90	80.0								80	80	100	100	90.0															
1440	1	4	90	100	70	86.7								100	90	90	90	92.5															
1543	1	6	90	80	100	90.0																		80	80	90	83.3						
1552	1	6	95	75	70	80.0																		60	70	70	66.7						
1420	1	6	70	40	80	63.3																		70	60	70	66.7						
1410	1	6	90	100	90	93.3																		90	90	90	90.0						
1442	1	6	90	90	90	90.0																		100	80	100	93.3						
1430	1	6	50	90	80	73.3																		50	80	75	68.3						
1544	1	7	80	80	70	76.7																						80	90	80	90	90	86.0
1431	1	7	100	100	70	90.0																						90	100	100	60	100	90.0
1502	2	3					95	100	100	98.3	98	100	99.0																				
1510	2	3					70	80	40	63.3	50	60	55.0																				
1522	2	4					75	60	70	68.3				80	85	80	85	82.5															
1443	2	4					70	60	70	66.7				80	80	70	80	77.5															
1554	2	5					20	40	40	33.3									30	20	70	80	50.0										
1415	2	5					90	80	90	86.7									100	100	100	100	100.0										
1525	2	7					80	90	40	70.0																		60	70	80	40	40	58 0
1417	3	4									90	60	75.0	90	70	100	70	82.5															
1534	3	5									100	80	90.0						90	90	80	100	90.0										
1506	3	5									70	60	65.0						70	80	50	60	65.0										
1403	3	5									60	85	72.5						100	90	90	90	92.5										
1422	3	6									70	60	65.0											60	70	70	66.7						
1530	3	7									90	90	90.0															8C	80	90	90	90	86.0
1555	3	7									20	10	15.0															70	90	80	80	90	82.0
1527	4	5												90	100	100	80	92.5	100	90	90	90	92.5										
1441	4	6												90	100	100	90	95.0						70	90	90	83.3						
1413	4	7												90	90	85	90	88.8										100	90	90	90	70	88.0
1448	4	7												90	60	80	90	80.0										60	70	60	80	90	72.0
1501	5	6																	90	95	95	95	93.8	85	90	85	86.7						
1425	5	6																	90	78	75	72	78.8	70	75	75	73.3						
1436	5	6																	100	100	100	100	100.0	90	90	100	93.3						
1541	5	7																	90	100	100	100	97.5					100	100	90	100	100	98.0
1513	6	7																						60	80	60	66.7	90	80	80	90	80	84.0
Means			84.1	83.2	81.8	83.0	71.4	72.9	64.3	69.5	73.8	69.5	71.7	87.8	83.9	89.4	86.1	86.8	86.0	84.3	85.0	88.7	86.0	73.8	79.6	81.3	78.2	81.1	85.6	83.3	80.0	83.3	82.7
Std. Dev.			13.93	17.65	10.79	9.24	24.62	20.59	25.07	20.34	25.25	25.87	24.37	6.67	13.18	11.30	8.58	6.28	21.71	23.91	16.33	13.87	16.58	15.24	9.64	12.99	11.18	15.37	11.30	11.18	18.71	18.71	11.53
EXPERTS			70	55	50	58.3	65	60	50	58.3	45	35	40.0	65	55	70	65	63.8	55	50	50	70	56.3	55	50	60	55.0	65	65	50	65	60	61.0

SCREENED.XLS, 7/22/96

Appendix BD. Treatment Group-Screened Second Evaluation Scores

Subject #	1st	2nd	1.1	1.2	1.3	1.0 Mean	2.1	2.2	2.3	2.0 Mean	3.1	3.2	3.0 Mean	4 1	4.2	4.3	4.4	4.0 Mean	5.1	5.2	5.3	5.4	5.0 Mean	6.1	6.2	6.3	6.0 Mean	7.1	7.2	7.3	7.4	7.5	7.0 Mean
2519	1	2	100	90	80	90.0	90	90	90	90.0																							
2533	1	2					80	60	50	63.3																							
2521	1	2	30	60	40	43.3	70	50	50	56.7																							
2416	1	2	30	40	40	36.7	50	40	40	43.3																							
2404	1	3	90	80	80	83.3					80	90	85.0																				
2419	1	4	90	80	80	83.3								90	80	80	90	85.0															
2402	1	4	100	100	80	93.3								100	100	100	80	95.0															
2509	1	5	100	90	100	96.7													100	90	90	90	92.5										
2412	1	5	80	80	70	76.7													70	50	60	60	60.0										
2407	1	6	70	90	70	76.7																		60	60	60	60.0						
2435	1	6	70	50	80	66.7																		50	60	50	53.3						
2408	1	7	80	90	100	90.0																						90	90	90	90	90	90.0
2545	2	3																															
2515	2	3					60	40	60	53.3	50	10	30.0																				
2505	2	5					50	60	60	56.7									80	80	70	60	72.5										
2537	2	5					80	60	70	70.0									80	60	80	70	72.5										
2553	2	5					60	60	70	63.3									80	80	70	80	77.5										
2427	2	7					40	40	70	50.0																		80	80	70	50	70	70.0
2508	3	4									90	70	80.0	60	60	90	80	72.5															
2449	3	4									60	50	55.0	90	80	80	80	82.5															
2405	3	4									60	50	55.0	80	70	90	80	80.0															
2507	3	5									80	40	60.0						70	40	60	70	60.0										
2550	3	5									40	40	40.0						70	40	60	60	57.5										
2411	3	6									70	90	80.0											80	80	90	83.3						
2535	3	7									50	70	60.0															90	90	80	90	70	84.0
2424	4	5												70	60	70	60	65.0	70	70	60	70	67.5										
2439	4	5												80	70	80	100	82.5	90	70	80	90	82.5										
2433	4	6												30	30	60	70	47.5						50	60	80	63.3						
2504	4	7												80	80	70	70	75.0										70	90	60	80	70	74.0
2546	4	7												70	80	90	90	82.5										80	80	70	50	70	70.0
2548	5	6																	60	80	50	50	60.0	40	30	40	36.7						
2429	5	6																	100	100	100	90	97.5	90	60	80	76.7						
2423	5	7																	60	90	60	90	75.0					30	70	60	50	50	52.0
2551	6	7																						30	60	30	40.0	90	100	70	70	60	78.0
2414	6	7																						60	60	70	63.3	90	80	80	70	80	80.0
2434	6	7																						60	50	60	56.7	60	80	60	20	70	58.0
Means			76.4	77.3	74.5	76.1	64.4	55.6	62.2	60.7	64.4	56.7	60.6	75.0	71.0	81.0	80.0	76.8	77.5	70.8	70.0	73.3	72.9	57.8	57.8	62.2	59.3	75.6	84.4	71.1	63.3	70.0	72.9
Std. Dev.			25.41	19.02	19.68	19.82	16.67	15.90	14.81	13.52	16.67	25.98	18.62	19.58	18.53	11.97	11.55	13.02	13.57	19.75	14.77	14.35	13.05	18.56	13.02	19.86	15.16	20.07	8.82	10.54	22.91	11.18	12.09
EXPERTS			70	55	50	58.3	65	60	50	58.3	45	35	40.0	65	55	70	65	63.8	55	50	50	70	56.3	55	50	60	55.0	65	65	50	65	60	61.0

CURRICULUM VITAE
GARRY D. COLEMAN

DATE OF BIRTH: June 24, 1961

ACADEMIC DEGREES:

1996	Ph.D., Industrial and Systems Engineering (Management Systems Engineering option), Virginia Polytechnic Institute and State University
1989	M.S., Industrial Engineering and Operations Research (Management Systems Engineering option), Virginia Polytechnic Institute and State University
1983	B.S., Mining Engineering, Virginia Polytechnic Institute and State University

CURRENT EMPLOYMENT:

Cunningham Fellow and Instructor of Industrial and Systems Engineering

OTHER RELATED EXPERIENCE:

1992-1993	Senior Research Associate, Virginia Quality & Productivity Center
1989-1993	Director of Business Development, Virginia Quality & Productivity Center
1988-1992	Research Associate, Virginia Productivity Center
1986-1988	Graduate Research Assistant, Virginia Productivity Center
1984-1986	Project Manager/Mining Engineer, Cannelton Industries, Inc.
1983-1984	Coal Miner, Chestnut Ridge Mining Co.
1978-1982	Cooperative Education Student and Summer Employee in the Coal Mining Industry, Biggs Branch Coal Co. and Southern Ohio Coal Co.

CONSULTING, PATENTS, ETC. OF LAST FIVE YEARS:

Consulting Projects:

Environmental Protection Agency Region III, ongoing 1993 - 1996
Value Engineering Office of the Westinghouse Hanford Company, situation appraisal, December 1992.
Ball Foundation, third party evaluator, 1991

STATE(S) IN WHICH REGISTERED:

Professional Engineer, Virginia, No. 020567

PRINCIPAL PUBLICATIONS OF LAST FIVE YEARS:

Coleman, G.D. and A. Clark, "Do Deming's 14 Points Apply to Research and Development Organizations?" Presentation accepted and paper under review for the proceedings of the 4th Industrial Engineering Research Conference, Nashville, TN, May 1995.

Coleman, G.D. and K.D. Black, "Applying a Grand Strategy Approach to TQM Implementation." Paper presented and published in the proceedings of the IIE 1993 International Industrial Engineering Conference, Los Angeles, May 1993.

Coleman, S.L., G.D. Coleman, and C.S. Johnston, "A Model for Managing and Measuring the Performance of the Training Function: Applying TQM Principles." Productivity and Quality Management Frontiers IV, refereed papers presented at the Fourth International Conference on Productivity and Quality Research, 1993, Miami, FL. Norcross, Georgia: Industrial Engineering and Management Press.

Coleman, G.D. and C.S. Johnston, "Implementing the Socio-Technical Systems Approach, Including Self-Managing Teams in a Start-Up Organization: An Interview with Bob Hoover, Plant Manager, Corning, Blacksburg." QPM: Quality and Productivity Management, Vol. 9, No. 4, 1992.

Coleman, G. D., D.S. Sink, and R. B. Horne, "Total Quality Leadership in Action in the U.S. Navy: A Case Study - System and Process Improvement in Ship Design, Acquisition, and Construction." Productivity and Quality with a Focus on Government, refereed papers presented at the First International Symposium on Productivity and Quality with a Focus on Government, 1992, Washington, D.C: IIE Press.

Coleman, G. D., and E. Van Aken, "Applying Small-Group Behavior Dynamics to Improve Action Team Performance," Employment Relations Today, Autumn 1991.

Sink, D.S. and G.D. Coleman, "A Strategic Management Process for the Organization of the Future." Productivity Measurement Handbook, The Management Innovations Group, Stamford, CT, 1991.

Coleman, G. D., and E. Van Aken, "Using Group Behavior Training to Improve Action Team Performance," Productivity and Quality Management Frontiers - III; refereed papers presented at the Third International Conference on Productivity and Quality Research, 1991, Miami, FL. Norcross, Georgia: IIE Press.

SCIENTIFIC AND PROFESSIONAL SOCIETIES OF WHICH A MEMBER:

Institute of Industrial Engineers

American Society for Quality Control

National Society of Professional Engineers

Virginia Society of Professional Engineers

IEEE Engineering Management Society

Society of Mining Engineers

HONORS AND AWARDS OF LAST FIVE YEARS:

Graduate School Cunningham Fellowship, Virginia Tech, 1993-96.

Service Award, Virginia Tech Chapter of Kappa Alpha Order, 1996.

Listed in Who's Who in Finance and Industry, 24th Edition, 1993.

Silver Award, 14th Annual American Homebrewers Association National Competition, 1992.

Silver Certificate, American Homebrewers Association Thirteenth Annual National Competition, 1991.

Outstanding Alumnus Award, Virginia Tech Chapter of Kappa Alpha Order, 1991.

SUBJECTS OR COURSES TAUGHT DURING THE LAST FIVE YEARS:

Second Summer Session 1996
First Summer Session 1996
Spring Semester 1996
Fall Semester 1995
Second Summer Session 1995

Course/Name:	ISE 2014	Engineering Economy			
Credit Hours:	2H, 2C	Sections:	1 per term	Type:	Lecture, Day

Spring Semester 1995
Spring Semester 1994

Course/Name:	ISE 4014	Compensation Management			
Credit Hours:	3H, 3C	Sections:	1 per term	Type:	Lecture, Day

Fall Semester 1994
First Summer Session 1994
Fall Semester 1993

Course/Name:	ISE 2024	Introduction to Industrial Engineering			
Credit Hours:	3H, 3C	Sections:	1 per term	Type:	Lecture, Day

OTHER ASSIGNED DUTIES AND SERVICE TO THE PROFESSION AND UNIVERSITY:

Co-Advisor to one Senior Design Team, 1994-95 and 1995-96.
President, Institute of Industrial Engineers (IIE) Chapter 81 - Western Virginia, 1995-96
Vice President, IIE Chapter 81 - Western Virginia, 1994-95
VEQTOR Chairman, IIE Chapter 81 - Western Virginia, 1993-94

Service on University Committees:
Member of Greek Life Advisory Council, 1993-96.
Member of Virginia Tech University Council, 1994-95.
Member of Virginia Tech Senior Vice President and Provost search committee, 1994.
Elected graduate member of Virginia Tech Student Budget Board, 1994-95.

SPECIFIC PROGRAMS TO IMPROVE TEACHING AND PROFESSIONAL COMPETENCE DURING LAST FIVE YEARS:

May 22, 1995	Attended ABET Retraining Session - IE Program Evaluators
November 14-18, 1994	Attended the Management Systems Engineering Center's (individual and organizational) Performance Improvement Bootcamp
Fall 1993	Completed EDCI 6644 - College Teaching, three credit graduate seminar. Have maintained a teaching portfolio as a result of this course.
June, 1992	Attended "Personality Types and Values" course (one day) taught by Leo McManus
July 1991, and January 1992	Attended "Instituting Dr. Deming's Methods for Management of Productivity and Quality" taught by Dr. W. E. Deming, Dr. W.W. Scherkenbach, Dr. B. Joiner, Dr. P Scholtes, Dr. E. Baker, and Mr. H. Hacqueboard (two day course); participated as helper/learner, July 1991, and attended January 1992.
May 1990 and September 1991	Attended "Quality, Productivity, and Competitive Position" taught by Dr. W. Edwards Deming (four day course)
October, 1991	Attended "Motivation and Management" course (one day) taught by Leo McManus
August, 1991	Attended "Guilt-Free Assertiveness" seminar (one day course)

PROJECT DIRECTION OF LAST FIVE YEARS:

"Developing a Corporate Grand Strategy System," SYSCON Corporation, 6/92-4/93, $31,000, co-principal investigator.

"Developing an Integrated Strategic Planning Process with TQM Principles," EPA Region III, 10/92-9/93, $80,000; funded in three increments, principal investigator.

"A Review of Studies Supporting or Refuting Dr. Deming's 14 Points," Naval Underwater Warfare Center, 10/92-5/93, $32,900, principal investigator.

"Strategic Planning for Va. Tech's Commercial Fish And Shell Fish Technology Programs," partially funded by the Virginia Tech Dept. of Food Science and cost shared by VPC, 6/92-9/92, $3,000 internal transfer, principal investigator.

"Strategic Planning/Total Quality Leadership Training Support," Naval Explosive Ordnance Disposal Technology Center, 1992-93, $63,000 follow on contract, co-principal investigator and project manager.

"Design and Development of a Strategic Management Process for the SUPSHIP Community," Naval Sea Systems Command, Supervisors of Shipbuilding, Construction, and Repair, 1991-1992, $25,000 initial contract and $99,000 follow on contract, primary deliverer and project manager.

"A Strategic Management Approach to Total Quality Leadership," Naval Sea Systems Command, Weapons and Combat Systems Directorate, 1991-92, $61,000 initial contract and $74,000 follow on contract, co-deliverer and project manager.

"Total Quality Management Services," Office of Personnel Management, selected as one of only twenty-four suppliers of TQM services for the Federal Supply Schedule, approved April 1990, renewed April 1991 and April 1992, project manager and principal point of contact. Author of proposal "Total Quality Management Implementation," OPM-RFP-91-02503, submitted November 1991, for the period of April 1993-1997, amount undetermined as budget is subject to task orders received (limit = $1,000,000 per task order). This proposal was for "re-certification" as a supplier on the Federal Government Supply Schedule for TQM services. Administration transferred to the General Services Administration during review. Technical proposal accepted without modification in 1992. Cost proposal accepted with modification in May, 1994. Primary delay due to third party protest of the award process.

"Continuous Improvement of the Ship Design, Acquisition, and Construction Process," Naval Sea Systems Command, Ship Engineering Directorate, 1990-1991, $150,000, project manager and facilitator.

"Research of Post-Classical Management Tools for Government Program Offices," special research grant proposal submitted to the U.S. Department of Energy by Virginia Productivity Center and Management Systems Laboratories (MSL), January 1990. Five year period of performance proposed, $9.8 million total funding. Grant funded for $1,000,000, Nov. 1990 to Nov. 1991. Lead author for VPC on joint VPC/MSL proposal team (three month effort). VPC project manager and interim grant administrator during first six months of grant.

Garry D. Coleman
7-23-96