ESTIMATING THIRD-PARTY EXAMINERS' SCORING STABILITY ON SELECTED APPLICATIONS TO THE TEXAS AWARD FOR PERFORMANCE EXCELLENCE

A Dissertation

by

BRANDI LYN PLUNKETT

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2006

Major Subject: Educational Human Resource Development

ESTIMATING THIRD-PARTY EXAMINERS' SCORING STABILITY ON SELECTED APPLICATIONS TO THE TEXAS AWARD FOR PERFORMANCE EXCELLENCE

A Dissertation

by

BRANDI LYN PLUNKETT

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Bryan R. Cole
	Toby Marshall Egan
Committee Members,	Homer Tolson
	Ben D. Welch
Head of Department,	Jim Scheurich

December 2006

Major Subject: Educational Human Resource Development

ABSTRACT

Estimating Third-Party Examiners' Scoring Stability on Selected Applications to the Texas Award for Performance Excellence. (December 2006)

Brandi Lyn Plunkett, B.S., Texas A&M University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Bryan R. Cole
Dr. Toby Marshall Egan

This study was an attempt to add to existing research by estimating the ability of third-party examiners to assess whether or not an organization successfully implemented strategies based on the criteria of the Texas Award for Performance Excellence (TAPE). The TAPE is given each year by the Quality Texas Foundation and recognizes organizations that demonstrate superior performance as it is defined by customer satisfaction and continuous improvement. The TAPE is a state-level award for quality that uses the same criteria as the Malcolm Baldrige National Quality Award for Performance Excellence.

This research was an analysis of the TAPE process at the level of examiners, also known as the Board of Examiners. The Board is made up of approximately 150 experienced professionals from several types of business sectors and is responsible for evaluating organizational self-assessments.

In this quantitative study, data were converted from the Quality Texas Foundation into a database. Because the set of the TAPE applicants included in the study

consisted of the entire population of TAPE applicants selected from 2001 to 2004, descriptive statistics were appropriate for producing informative data that could be analyzed for variation and stability in the scoring process. Exploration of patterns in descriptive statistics and multivariate analysis of variance were the primary tools used in this particular study along with Cronbach's Alpha as an indicator of reliability.

Since scoring for the TAPE is based on an individual examiner's best subjective assessment, it was impossible to have one objective score against which all the other scores could be measured. The team consensus score was therefore used as the true score for measurement.

Establishing reliability of examiners' scores was a problem due to the fact that organizations and teams did not repeat. Results from the study led to the conclusion that there was insufficient evidence to make a determination on what influences examiners' scoring consistency. More data will need to be collected in such a way so as to make it possible to identify that impact consistency of examiner scores.

DEDICATION

This work is dedicated to my late grandfather, Joe Bob Plunkett,

my grandmother, Helen, my brother, Joe,

and my parents, Bob and Judy Plunkett.

This was a long, hard journey made bearable by your constant support.

You provided inspiration by your example, unwavering confidence in me,

and unconditional love, all my life.

Thank you, and I love you.

ACKNOWLEDGMENTS

What a time of challenging growth this has been! In truth, the enormity of the completion of this period of life has not yet hit home. I was recently asked to describe how I've changed since entering graduate school in 2001 for M.S. and Ph.D. degrees. The best answer I could give at the time was that the world has become a larger place for me. Doors of awareness have been opened, and I have a broader perspective from which to generate better questions—not necessarily answers. I have learned that asking the right questions and being open to the replies is much more important than always having an answer, for it is in the questioning of things that one finds next steps.

Many were the times that I had questions along this journey, and there were so many wonderful people who were there to provide guidance, answers, more questions, ideas, friendship, love, and support. At times like this, when I find myself consumed with humility and gratitude, the thought of trying to name each person and his or her particular influence on my life seems almost overwhelming. I know that who I am today is the sum total of all the experiences I have had, experiences shaped by the people in my life. What follows is an attempt to express sincere appreciation to those who have worked with me on this endeavor.

First, my thanks to Dr. Bryan Cole, who saw something in me a long time ago when I was teaching kindergarten in the Leander ISD. Thank you for bringing me the application to graduate school, for giving me a job as your graduate assistant, and for opening my mind to the world of high performance and continuous improvement.

You provided the most challenging experiences I faced in graduate school and the foundation upon which I have built my professional philosophy. Those challenges pushed me far beyond anything I ever thought I could do. I didn't give up because I had you as an example, and because I couldn't live with the thought of letting you down. I watched you work harder than anyone I've ever seen during the years you served as head of our department and led the Bonfire Commission simultaneously, and I learned what it means to be truly committed to excellence. Second only to my own father and grandfather, you are a hero in my eyes. Thank you for helping to shape who I am today and for caring about who I am. Your opinion means the world to me, and it always will.

Second on the list, but not in the heart—I want to express my deepest appreciation to Dr. Toby Egan. I have said this over and over again; I would not have made it to the end of this journey were it not for you. I could talk about all that you have done for me as a professor and co-chair, but it was your personal encouragement that kept me moving forward. You have been there for me, not only as an advisor and co-chair, but as a counselor and friend. You listened to me when I was in the depths of despair over lost identity, lost jobs, lost family and pets, lost purpose, even lost or corrupted dissertation chapters, and you were always, always there to listen and help me find my way. You know better than anyone what this journey has been like for me and what is has cost. You have believed in me when I didn't or wouldn't believe in myself. Having spent a long time in higher education, I find that I have even more respect for those professors who are truly teachers; they are few and far between. You are a true teacher, Toby. You take the time to really make a difference in the lives of

your students and, at the same time, you are a brilliant professor and researcher. My admiration for your intellect, your heart, and your commitment to living your values is immeasurable, as is my deep and profound gratitude for everything you've done for me.

Selfless service is a rare quality found only in very special individuals of the highest character and integrity. I have been blessed to know several people like that, two of whom are among the most generous men I know, Dr. Homer Tolson and Mr. Bill Ashworth. Dr. Tolson, your incredible patience and willingness to walk me through the statistical labyrinth of this project was truly selfless. Thank you for your open door, for meeting me with a smile, and for your calming assurance that I would find my way to the end of this journey. Thank you for doing whatever it took to help me understand and work through the most difficult elements of my study. You are a remarkable man and a true teacher.

Mr. Bill Ashworth, your commitment to me and my success has been incredibly selfless. Thank you for the countless late hours you spent helping shape and reshape the format of this work. I don't know anyone who would work so hard for the benefit of someone else. This project had some incredibly short timelines, but you made it happen. You worked tirelessly and with tenacity, and I am just so blessed to have had you to lean on. Thank you for being on my team, for standing behind me through this journey and for your high character and ethics and your commitment to this project. You are truly a blessing.

There are many friends who, simply by being a part of my life, have provided incredible support and made this journey a richer one. Thank you to Tom Clark with

whom I share a unique and rewarding friendship. The conversations we had and the synergy we shared as we navigated our way through what seemed like an obstacle course at times made my graduate experience rewarding and were among the brightest spots in my learning. Even though we don't get to talk very often, you will always have a very special place in my heart. The "Mixed Nuts"—Dr. Manda Rosser, Dr. Matt Upton, and Ms. Dorian Martin—you made learning fun and provided a sense of family during the last few years of this trip. Thank you for walking the long road with me.

Finally, my Aggie friends—Marianna, Stroh, Hilary, Kim, Julie, Lisa, Mary Olga, John, and so many others—thank you for being there to encourage and listen and for just being interested. Many of you have said you felt like you were going through this with me—and you were. You were there to play and relax when I needed to take a step back and you were there to hold me up when times were tough. You each have added to my life and blessed it beyond words. I can't imagine what the past five and a half years would have been like without you in my corner, and I can't imagine what the future would be like without you.

As I said before, who I am today is the sum total of all the experiences I have had, experiences shaped by the people in my life. Thanks to all of you for your investment of time and talent, for your guidance, your support and for making this challenging journey a marvelous experience!

TABLE OF CONTENTS

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Quality Awards

Quality Management and the Malcolm Baldrige Award for Performance Excellence have been important terms used in business lexicon for several decades. The Baldrige Award has had a profound impact on modern American business and has impacted business standards around the world (Malcolm Baldrige National Quality Award, 1998; Vokurka, 2001; Vokurka, Stadling & Brazeal, 2000). Today, there are at least 75 countries that have national awards (APQC's Knowledge Sharing Network (n.d.); United Nations Economic Commission for Europe (2004). Quality Digest (2005), lists at least 32 states within the United States having state-level awards for performance excellence that are modeled after the national award. However, the reported numbers vary. Other sites such as the Network for Excellence (n.d), referred to by the National Institute for Standards and Technology, lists 82 quality awards around the country. Another source listed 42 states with quality award programs (Network for Excellence, n.d.). Regardless of the exact number, it is clear that quality awards, including state, national and international, are a growing phenom-enon.

The style and format for this dissertation follow that of the *Journal of Educational Research*.

Lack of Research

Most of these awards, including the Baldrige, are built on a foundational belief that third-party examiners responsible for assessing organizations can consistently and accurately determine, based on the organization's self-assessment, which organizations meet or exceed the established criteria setting it apart from other applicants. Winning organizations receive recognition as leaders in achieving performance excellence. Once they receive the tremendous accolades and publicity that come with winning the award, they are hence forth regarded as exemplars of how to implement quality principles. One troubling fact exists, however: Empirical evidence validating the ability of third-party examiners' to accurately assess organizations is remarkably scarce (Coleman, Koelling, & Geller, 2001; Conti, 1994).

For those in the Human Resource Development (HRD) field, a lack of confidence in the accuracy of an organizational assessment based on the Baldrige or any other quality award produces questions about how these examiners are trained. Those who specialize in the HRD function of training and development of quality award examiners may be able to use this study to inform their training techniques and activities.

Identification of the Study

This study was an attempt to add to the existing research by looking at the ability of third-party examiners to assess whether or not an organization successfully implemented strategies based on the criteria of the Texas Award for Performance Excellence. The Texas Award for Performance Excellence (TAPE) is given each year

by the Quality Texas Foundation and recognizes organizations that demonstrate superior performance as it is defined by customer satisfaction and continuous improvement (About Quality Texas, n.d.). The TAPE is a state level award for quality that uses the same criteria as the Malcolm Baldrige National Quality Award for Performance Excellence.

History of Quality and the Malcolm Baldrige National Quality Award

"Quality" as a descriptor of management philosophy has existed in the U.S. business vocabulary for decades. It became particularly prevalent in 1980 when a television documentary entitled *If Japan Can…Why Can't We* aired on American television. This documentary shook the U.S. psyche as it spotlighted the success of Japanese industry and exposed the fact that U.S. business was slipping in the world economy. The documentary introduced us to W. Edwards Deming and his total quality approach that helped Japan turn their economy around and dominate the global market economy. Since 1980, American business has undergone a major paradigm shift from quality defined by lowest cost production to quality defined by customer satisfaction (Hoyer & Hoyer, 2001).

During the early years of the quality revolution, many American businesses did not know where or how to begin to change their organizations and still others did not recognize the significance for change (Vokurka, 2001). Leaders in industry and government worked together to develop the Malcolm Baldrige National Quality Award (MBNQA) which was signed into existence on August 20, 1987 through

Public Law 100-107 and placed under the management of the National Institute for Standards and Technology (NIST) (NIST, 1998).

The purpose of the MBNQA is to "… recognize U.S. organizations for their achievements in quality and business performance and to raise awareness about the importance of quality and performance excellence as a competitive advantage" (NIST, 1998). The thought behind this award was that quality management was the best strategy to create benchmarks to which U.S. organizations should aspire in their quest for world class quality (Vokurka, 2001).

Texas Award for Performance Excellence

The MBNQA has not only set benchmarks for quality in business, it has also set benchmarks for other quality award programs both internationally as well as within the U.S. One such program is the Texas Award for Performance Excellence (TAPE). The TAPE, which is patterned after the MBNQA, was established in the early 1990s as a result of the combined efforts of the governor's office, the Texas Department of Commerce, and Texas businesses. In 1994, Quality Texas was established as the independent non-profit administrative corporation of the TAPE (About Quality Texas, n.d.).

The TAPE is made up of the same categories and evaluation procedures as the MBNQA. It is an evaluation of organizations from six sectors of business including education, health care, public organizations, small organizations, service, and manufacturing. Each of these organizational sectors has its own version of judging criteria tailored to their unique needs, and all are judged on the same seven categories which

include leadership, strategic planning, customer and market focus, information and analysis, human resource focus, process management, and results.

The starting point of both the Baldrige and the TAPE process is an organizational self-assessment followed by a third-party examiner assessment and, in some cases, a site visit by the examiners. For the TAPE process, there are three levels of application from which organizations can choose based on their level of experience with quality including Self-Assessment Level, Progress Level and Award Level. The highest level, which is the focus of this study, is the Award Level Process. The entire set of seven performance criteria used for the Baldrige Award is utilized in the Award Level. This level is mainly used by those organizations that have been using the principles and practices of performance excellence for a significant period of time (Quality Texas Foundation, 2005).

Organizations applying at this level must begin by submitting an application (50-page limit) and an organizational profile that gives examiners an overview of the particular organization. After the application is submitted, a team of examiners assesses the organization based solely on the application and profile. They first examine the documents individually, fill out a comprehensive scoring booklet and then meet as a team for the purpose of coming to consensus on every item of the scoring criteria. Once the examiner team comes to consensus, a Panel of Judges determines which applicants will be awarded site visits. Applicants who do not receive site visits will receive a detailed feedback report. Applicants who do receive a site visit, which is simply a visit to the organization for the purpose of verifying the application, will receive a feedback report after the visit is completed and the

examiner team has had more time to review their findings. This feedback report is an invaluable resource for each organization as a continuous improvement tool. A final report of all evaluations of the applicants receiving site visits is conducted by the Panel of Judges who develop recommendations and pass them to the Board of Directors. The Board of Directors makes the final decision on who should receive the TAPE.

Research

This research was an analysis of the TAPE process at the level of examiners. With both the TAPE and MBNQA, this volunteer group of examiners is known as the Board of Examiners. The Board is made up of experienced professionals from the private, public, education, and health care sectors and is selected by a governing board called the Board of Overseers through an application process. The Board of Examiners consists of approximately 150 members and is responsible for evaluating applications, preparing feedback reports, and conducting site visits.

Examiner teams are a heterogeneous mix of new, returning, and senior examiners made up of men and women of various ages and experience levels. New examiners are individuals who are working with the TAPE for the first time; returning examiners have had one or two years of experience; and senior examiners are considered veterans not only in the TAPE process, but also in the field of quality.

All examiners, regardless of experience, must attend a three-day examiner training session every year. Prior to attending training, examiners must conduct an examination of a faux organization to be used as a case study during the three days of

training. During the training, examiners work in teams to analyze the case study and learn about each item within each category of the TAPE, how to write non-prescriptive feedback, and how to identify strengths and opportunities for improvement regarding various criteria items. They also learn how to score items and come to consensus with other examiners on their team. This process of using outside examiners to assess organizations is known as "third-party assessment" and is the focus of the research.

An initial review of the literature on the effects of training and performance on quality award examiners yielded few results (Berquist, 1996; Coleman, 1996; Coleman et al., 2001; Coleman, Van Aken, & Shen, 2002). Concern for the accuracy and variability of examiners' scores, however, has been mentioned in the literature beginning with several papers submitted to the First European Forum on Quality Self-Assessment in 1994 (Conti, 1994; Fuchs & Stuntebeck, 1994; Jernberg. Lindstrom, & Chocron, 1994; Martellani, 1994). Not enough research has been conducted on the factors affecting accuracy of examiners (Conti, 1994).

Garry Coleman of the University of Tennessee has conducted much of the recent inquiry into the reliability of examiner scoring. His 1996 dissertation was an estimation of the impact of third-party examiner training on the scoring of organizational self-assessments by conducting an experiment using the 1995 Baldrige case study and 81 graduate students. The goal was to estimate the impact of training and explore the relationship between examiner characteristics and score accuracy (Coleman, 1996). Coleman has written other papers on the training and scoring accuracy of self-assessments looking at several variables including accuracy indices,

types of training, and interrater reliability. Keinath and Gorski (1999) and Sienknecht (1999) also studied the interrater reliability of examiner scoring while van der Wiele, Williams, Kolb, & Dale (1995) studied the variance of examiners' scores (as cited in Coleman, 2000).

While scholarly research has not produced definitive results on how to reduce third-party examiner variation, steps have been taken by the quality award organizations to reduce the potential for examiner error (Coleman et al. 2000). Organizations like the MBNQA and TAPE build heterogeneous pools of quality experts from which to draw teams of examiners who will be assigned to various applicants. Examiners for the TAPE are required to assess a case study of an organizational self-assessment and then attend an intense three-day training with other examiners before being assigned to a team. While training and selection criteria are likely helpful in addressing the potential for error, not enough is known about the effects of training on examiners to assess if it is truly effective.

Statement of the Problem

Quality award assessments such as the Baldrige Award and the Texas Award for Performance Excellence are widely accepted as an efficient and effective way to measure organizational performance (DeBaylo, 1999). Findings of third-party reviews are increasingly used for decision-making and change initiatives (Coleman, Koelling, & Geller, 2000; Coleman et al., 2002); however, there is little research that clearly establishes that the process of third-party examination is accurate and objective. While concern about the stability of examiner scoring has been voiced in

the past (Conti, 1994; Fuchs & Stuntebeck, 1994; Martellani, 1994), there is a relative lack of scholarly analysis of third-party examiners and their ability to objectively assess organizational performance (Coleman, Koelling, & Geller, 2000; Coleman et al., 2002; Conti, 1994). Rather, business and industry operate on the assumption that the success of organizations who receive recognition through quality awards is due to their overall business performance, showing little concern for potential error in the assessment process. Given the lack of research, it cannot be concluded that organizations are judged consistently over time and across the sectors and categories of the quality awards. Therefore, it is important that more research on examiner training and scoring be conducted.

Purpose of the Study

The purpose of this study was to determine the scoring stability of third-party examiners who were assessing organizational performance for the Texas Award for Performance Excellence.

Research Questions

1. Is the mean of the deviations of individual total scores from team total consensus scores equal to zero?
2. Is the mean of the deviations of individual item scores from team item consensus scores equal to zero?
3. Do item deviation scores vary across the following classifications:
    a. Levels of Examiner Experience

b. Sector

c. Levels of Self-Assessment

d. Levels of Team Experience

Operational Definitions

The terms below were used in this research study based on the corresponding definitions.

Scoring Stability—refers to the consistency in item scores across several examiners for the Texas Award for Performance Excellence.

Levels of Team Experience—3 levels of team experience (Senior, Average, and New) were developed including teams with 51% or more senior examiners, teams with 51% or more new examiners, and teams which were 50% new examiners and 50% senior examiners. Returning examiners were combined with senior examiners for this research study.

Texas Award for Performance Excellence (TAPE*)*—the non-profit organization in the State of Texas that assesses organizational performance based on the quality philosophy and seven categories used in the Malcolm Baldrige National Award for Quality (Quality Texas Foundation, 2005).

Malcolm Baldrige National Quality Award (MBNQA)—the award program governed by the National Institute for Science and Technology in the United States that assesses and recognizes organizational performance based on the approach, deployment, and business results of quality principles. The MBNQA is the leading model for quality awards around the world (NIST, 1998).

Third-Party Examiner—an individual who has completed the TAPE training and has read and assessed an organization's application to the TAPE process (Quality Texas Foundation, 2005).

Category—one of seven areas addressed on the organizational assessment. Categories for the Baldrige and TAPE assessment include Leadership, Strategic Planning, Customer and Market Focus, Information and Analysis, Human Resource Focus, Process Management, and Results (Quality Texas Foundation, 2005).

Embedded Item—sub-categories within a category

Sector—the differentiation between various types of organizations. Sector titles for the Baldrige and TAPE include Small Organizations, Manufacturing, Education, Health Care, Public Business, and Service.

Self-Assessment Score—the score an examiner gives him/herself to illustrate the level of confidence in his/her ability to assess an organization

Individual Total Score—the final score given by one examiner to an organization prior to the team consensus meeting.

Total Consensus Score—the score arrived at through consensus of all examiners who assessed a particular organization. For the purpose of this study, the total consensus score were used as the "true score" against which individual scores will be measured.

Individual Total Deviation Score—the score produced by subtracting the total consensus score from the individual total score. [ITDS =ITS – TCS]

Team Mean Deviation Score—the mean of the individual total deviation scores for a team. TMDS = [(ITS-TCS_1) + (ITS-TCS_2) + (ITS-TCS_3)…..+ (ITS-TCS_n)] / n

Individual Item Score—the score given by one examiner for each of the 17 embedded items in an organizational assessment. This score is given prior to the team consensus meeting.

Team Item Consensus Score—the score given to an item by the team of examiners as a result of a consensus meeting.

Item Deviation Score—the score produced by subtracting the team item consensus score from an individual item score. (IDS =IIS – TICS)

Item Mean Deviation Scores—the mean of the item deviation scores for a team. There are 17 item mean deviation scores for each team.

IMDS = [(IIS-$TICS_1$) + (IIS-$TICS_2$) + (IIS-$TICS_3$)…..+ (IIS-$TICS_n$)] / n

Assumptions

The following assumptions were applied to this research.

1. The statistical analyses will accurately reflect the consistency in examiners' scoring and the effects of examiner experience, team experience, sector, and self-assessment on scores.
2. The interpretation of the data collected will accurately reflect what it was intended to reflect.

Limitations

The following limitations were applied to this research.

1. The scope of this study is limited to the four years of data collected on the Texas Award for Performance Excellence.
2. The consensus score is used as the "true score" against which other scores are compared.
3. The makeup of examiners and the ratio of experience levels are not consistent across the four years.
4. The training material and activities varied each year according to the changes in the TAPE criteria and the individual staff members who delivered training.
5. Small changes were occasionally made to the award criteria and therefore are slightly different in some categories from year to year.
6. Each team rates a different organization each year.
7. Organizations applying for the TAPE are at different experience levels of quality management and organizational self-assessment.
8. Findings from this study may not be generalized to any other quality award.

Significance of the Study

Since 1991, state and local quality award programs, most modeled after the Baldrige program, have grown from fewer than 10 programs to more than 80 in at least 41 states (Network for Excellence, n.d.; NIST, 1998). Internationally, approximately 90 quality programs are operating awards (APQC's Knowledge Sharing Network, n.d.; United Nations Economic Commission for Europe, 2004).

Since 1988, more than 1000 applications have been submitted for the Baldrige Award from a variety of types and sizes of organizations (NIST, 1998). The TAPE is almost identical to the Baldrige Award in its assessment process. Consequently, results from this study have the potential to promote further inquiry in the area of training for improved organizational performance.

Findings from this study will inform the Quality Texas Foundation regarding the stability of examiner scoring for the Texas Award for Performance Excellence Program. Additionally, the results of this study may provide insight into what influences examiners' scores leading to improved examiner training. Improved training could result in increased accuracy and objectivity where examiners are able to consistently identify strengths and opportunities for improvement within organizations, thereby increasing the reliability of the assessment process. When a level of stability can be established for examiners' scores on assessments, there can be more certainty that differences in organizational assessments are not a function of examiner differences, and that organizations applying for the TAPE are evaluated in a consistent manner.

Those responsible for training within organizations, specifically Human Resource Development (HRD) professionals, must continually seek to improve their planning and development of training programs. Results of this study have the potential to inform HRD practitioners about what impacts the way third-party examiners of the Texas Award for Performance Excellence view organizations. Consequently, HRD professionals may be able to help leaders and managers of Texas organizations better

understand how to produce clear and effective organizational self-assessment documents to gain accurate and reliable examinations.

CHAPTER II

REVIEW OF LITERATURE

Evolution of Quality from Quality Control to Continuous Improvement

The concept of "quality" can be discussed as a discipline, a philosophy, a theory or practice. It is sometimes described as an abstract concept that is defined "in the eye of the beholder," and other times described as an objective and measurable product outcome. The perceptions of quality and definitions of quality vary depending on the context within which the describer resides. Garvin (1988) pointed out that scholars from philosophy, economics, marketing, and operations management have all discussed quality in their respective literature. As a result, quality can be defined from several different perspectives including product-based, user-based, manufacturing-based, and value-based (Garvin, 1988). While demand for quality has been woven into the fabric of human nature for centuries, the establishment of quality standards and the attempt to measure quality only began as recently as the 20th century. The role of quality has changed for manufacturers and other organizations from inspection at the end of a process to quality assurance in the design of the process (Pryor, 1998). Today, the demand for quality by consumers is present in almost every product and service; it is a common term in the national and global marketplace and refers to the degree of customer satisfaction based on several different variables (Hoyer & Hoyer, 2001).

Over the past few decades, the world has witnessed a growing emphasis on quality. There seems, however, to be almost as many definitions of quality as there are agents trying to define it. Several notable authors have established themselves as

well-known quality experts, yet even among the experts, there is a poignant lack of consensus on the definition of quality. Some of these experts include Philip Crosby, W. Edwards Deming, Armand Feigenbaum, Kaoru Ishikawa, Joseph Juran, and Walter Shewart (Hoyer & Hoyer, 2001). In comparing the definitions of several of these quality experts, Hoyer and Hoyer found that definitions tended to have two perspectives in their approach to determine a definition. One type of definition stemmed from the perspective that quality is defined by measurable characteristics that, "satisfy a fixed set of specifications that are usually numerically defined" (2001, p. 54). A second type of definition stemmed from a more complicated view that quality is not fixed. Rather it is defined by customer expectations and satisfaction and is ever changing. The following paragraphs summarize the mentioned quality experts' perspectives on how the term "quality" should be interpreted.

Philip Crosby. One of Crosby's main issues as he attempted to define quality was that many use quality in a way that makes it a relative term whereby the meaning changes with the perspective of the user. As a result, he focused on the idea that the ability to define quality lies in the knowledge of specific product or service requirements stated in numerical terms (Hoyer & Hoyer, 2001). For business, Crosby states that, "Requirements must be clearly stated so they cannot be misunderstood. The nonconformance detected is the absence of quality. Quality problems become nonconformance problems, and quality becomes definable" (Crosby, 1979, p. 7).

W. Edward Deming. Deming's famous book, *Out of Crisis,* has been a guide for many who seek to understand quality improvement. Deming was one of the first to gain recognition as a quality "guru" for his work that transformed Japanese industry

in the post WWII era. He was so instrumental in the facilitation of Japan's recovery from postwar economic devastation of the war that they named their national quality award after him. The Deming Prize is the Japanese equivalent to the United States' Baldrige Award. Despite, or perhaps due to his perspective on quality improvement, Deming resisted an all-encompassing definition of quality. Rather, he discussed (in his book *Out of Crisis*) that ultimately, the definition of quality resided with the customer in terms of customer satisfaction. He also pointed out, however, that quality is a multi-dimensional notion with multiple characteristics and players, all of whom have a different perspective on what defines quality (Hoyer & Hoyer, 2001). For example, Deming illustrated that a production worker, a production manager and a customer will each define quality from the perspective of their point of interaction with a particular product or service (Deming, 1986). He further pointed out that, from the customer's perspective, needs and relative return on investment measures change with time, making an attempt at measuring one particular characteristic for quality difficult, if not impossible. Ultimately, Deming suggested that customer satisfaction, with its ever-changing requirements, is the best way to determine quality.

Armand Feigenbaum. Like Deming, Feigenbaum described quality as being dynamic and determined by customer satisfaction. He believed that quality is determined by customer satisfaction which is based upon the actual experience of the customer with the product or service. He asserted that satisfaction is measured against the customer's requirements whether they are stated or not, conscious or unconscious, operationally defined or subjective. Feingebaum believeed that quality, as defined by

customer satisfaction, is a moving target in a competitive market like that of today's (Feigenbaum, 1983).

Kaoru Ishikawa. Like Deming and Feigenbaum, Ishikawa defined quality from the perspective of customer satisfaction. In his book, *What is Total Quality Control? The Japanese Way*, Ishikawa asserted that, "We engage in quality control in order to manufacture products with the quality which can satisfy the requirements of consumers. The mere fact of meeting national standards or specifications is not the answer. It is simply insufficient" (Ishikawa, p. 44). He looked at measurable standards in terms of the national standards of the Japanese Industrial Standards or the International Organization for Standardization. This was a variation from the points made by American quality gurus, but like the American gurus, he situated the essential determination of quality with the consumer, adding that the price of a product (in terms of perceived value) is also an important factor in determining quality.

Joseph Juran. "Fitness for use" is the phrase Juran used in his attempt at an all encompassing definition of quality. He admitted that a practical definition is probably not possible. Recognizing that the term quality has many different meanings, he idenfified the two most dominantly used meanings; one that relates to customer satisfaction and one that refers to freedom from defect (Juran, 1989). Yet he offered "fitness for use" as a definition to, "… standardize on a short definition of the word 'quality'" (1989, p. 2). This definition is somewhat vague. The reader is left with the job of trying to determine the scope of the word *fitness* and the exact connotation of

the word *use*. Given that the reader will likely determine the meaning based on his or her own situation, the definition provides little clarity.

Walter Shewart. As early as the 1920s Shewart talked about the two different ways of looking at quality – objective and subjective. He discussed the subjective measure of quality as being of particular interest to those on the commercial side of the house because quality encompasses four different types of value, including use, cost, esteem and exchange (Shewart, 1931). In addition, Shewart understood and articulated the importance of objectively defining quality by using statistics to measure quality standards in terms of a fixed, achievable state.

As a paradigm, quality has evolved from an idea of post-production inspection to one that is proactive and strategic. Garvin (1988) describes four major *eras* of the quality evolution including the inspection era, the statistical quality control era, the quality assurance era, and ending with the strategic management era of today. The Inspection Era can be traced back to the days of artisans and skilled craftsmen of the eighteenth century. However, the scope of this section will begin with the 1900s and Frederick Taylor, the father of Scientific Management. He gave legitimacy to the idea that post-production inspection was a task that lead to better quality and described it as an activity that was the responsibility of supervisors in effective management (Garvin, 1988; Juran, 1995; Lindsay & Petrick, 1997). G. S. Radford, who in 1922 wrote *The Control of Quality in Manufacturing*, was the first person to discuss quality as a management responsibility and the need for cross-functional discussions, also stressed inspection as the primary function of quality production (Garvin, 1988).

As the next decade approached, a new definition for quality took shape with research from the Bell Telephone Laboratories and the publishing of W. A. Shewhart's book, *Economic Control of Quality of Manufactured Product*. According to Garvin (1988), this was the birth of the Statistical Quality Control Era of the 1930s and 40s. With the publishing of his book, Shewart gave "scientific footing" (Garvin, 1988, p. 6) to the discipline of quality for the first time in its history. He was the first to establish that variation would always exist in manufacturing due to such things as raw materials, varying skill levels of individuals, equipment and machinery. He demonstrated that variation could be understood and reduced through statistical analysis (Lindsay & Petrick, 1997). Shewart, along with other scientists at the Bell Laboratories, including Joseph Juran, developed process control and sampling techniques that would lead to the improvement of telephone equipment and service (Garvin, 1988; Juran, 1995; Lindsay & Petrick, 1997). Juran believed that sampling techniques and control charts were key elements in the development of quality control (Juran, 1991).

Eventually, statisticians would help the War Department in developing the concept of "acceptable quality levels" (AQL) which would tell manufacturers of arms and ammunition the minimum level of quality that would still be considered acceptable. This was followed by a revised inspection process that allowed for increased production of war materials at improved levels of quality (Garvin, 1988; Juran, 1995). This new technique of using statistics to control variation and improve inspection was now used in training for other branches of industry. According to Garvin (1998, p. 11), "By the end of the war, institutions in twenty-five states were

involved. A total of 8,000 people were trained in courses ranging from one-day executive programs to intensive eight-day seminars for engineers, inspectors, and other quality control practitioners." Many of the students got together to form groups that eventually led to the creation, in 1945, of the Society of Quality Engineers and later the American Society for Quality Control (ASQC) (Garvin, 1988; Juran, 1995). ASQC put out the first U.S. journal on the subject of quality control which is now called *Quality Progress*. The ASQ and *Quality Progress* magazine are the leading sources of information on quality control today.

For the rest of the 40s and into the 50s, Statistical Quality Control was the focus or discipline of quality; however, it was limited in scope as it only addressed what happened on the factory floor and only used statistical methods of analysis. As companies began to question exactly how costly defects were when products were not produced correctly, managers found themselves without means to calculate an answer. Joseph Juran addressed this concern with his 1951 edition of the *Quality Control Handbook* where he discussed economics of quality and the notion of avoidable and unavoidable costs (Lindsay & Petrick, 1997). This was the beginning of the Quality Assurance Era (Garvin, 1988).

Armand Feigenbaum added to the discussion by introducing the notion of "total quality control" (Lindsay & Petrick, 1997). He suggested that "quality is everyone's job," and that there were three basic stages that all new products went through, "… new design control, incoming material control, and product or shop floor control," (Garvin, 1988, p. 13). Cooperation from all departments involved in the production process was necessary to achieve an acceptable level of quality, hence the "total"

element of total quality control. Both Juran and Feigenbaum also introduced the idea that quality control was more than statistical in nature. They both discussed in their books, the need for new product development, vendor selection, and customer service to be added to the quality system (Garvin, 1988; Pryor, 1998).

Reliability Engineering, which was a quality term referring to the reliability of a product when it was used in the field over a long period of time, was also emerging (Garvin, 1988; Juran, 1995). According to Juran (1995, p. 561), "Reliability evolved to develop tools and procedures that would contribute to reducing field failures." Similar to total quality control, reliability engineering used engineering and attention to quality throughout the quality system to prevent defects (Garvin, 1988).

Zero Defects was the last movement to develop during the Quality Assurance Era (Garvin, 1988). It began at the Martin Company in the early sixties when a defect-free missile was promised and delivered to the U.S. Army's missile command. From this single event, management at Martin concluded that, "The reason behind the lack of perfection [in the past] was simply that perfection had not been expected. The one time management demanded perfection, it happened!" (Halpin as cited in Garvin, 1988). This sparked a new emphasis on workers' motivation and awareness and a new debate on the merits of acceptable quality levels (AQL) and the "perfect quality" of which Philip Crosby wrote in his book *Quality is Free* (Garvin, 1988).

From the early 1900s through the end of the Quality Assurance Era in the 1960s the paradigm of quality took a defensive stance of preventing defects. Into the 1970s and 80s, however, a more strategic and proactive approach to quality was being pursued. The perception of quality was changing from something that could hurt a

company if not tended to, to something that could actually make the company more competitive (Garvin, 1988; Pryor, 1998).

Since the NBC video entitled “If Japan Can, Why Can’t We” was aired in 1980 introducing W. Edwards Deming to American managers and highlighting the beginning of the Quality Revolution, strategic quality management has become a major source of study and discussion in the world of business. This video told how the Japanese had come to dominate the auto and electronics markets by following Deming’s advice to continually improve their processes and think of manufacturing as a system, not as pieces and parts (Garvin, 1988). Deming had helped the Japanese rebuild their economy which was devastated during WWII with his lectures on basic principles of statistical control of quality in the 1950s. The Japanese not only embraced Deming and his teachings, they also named their quality award, the most prestigious award given in Japan, after him; the quality prize in Japan is known as the Deming Prize (Lindsay & Petrick, 1997).

Most credit the NBC video with the increased attention and heightened sense of urgency resulting in the popularity of statistical process control (Juran, 1995) and continuous improvement leading to the development of the Malcolm Baldrige National Quality Award (MBNQA) in the United States. There were, however, other external forces responsible for the awakening of companies to the link between low quality and loss of profitability.

The primary force was, as the video pointed out, the increase of foreign competition, namely from the Japanese. There were also reports, however, from surveys taken between 1973 and 1983 that showed U.S. consumers were losing

confidence in the quality and reliability of American made products, and that they believed product quality had declined in the past five years (Garvin, 1988). Additionally, the 1970s and 80s saw more product liability suits and governmental pressures, such as an increased number of recalls from the National Highway Traffic Safety Administration, the Environmental Protection Agency, and the Consumer Product Safety Commission and new laws enacted, such as the lemon law designed to protect consumers from poor quality products (Garvin, 1988; Juran, 1995). All of these issues were costly to manufacturers and produced a sense of urgency to find a solution to the quality problem. Executive level leadership was now beginning to realize that solutions could no longer be relegated to middle management; they began to link quality to profitability, define it from the customer's perspective, and use it for strategic planning purposes (Garvin, 1988).

The added dimension making the Strategic Quality Management era unique was it's broadened perspective on what defined quality. Upper-level managers, who were now involved in the quality process, realized that focusing on eliminating defects and quality assurance was too narrow to be competitive. Quality had to be defined in terms of customer needs. This new perspective required that quality be defined both from a comparative and relative standpoint. Comparative quality relates to the performance of an organization with respect to its competitors while relative quality relates to how the organization's customers view the quality and value of the products and services. Quality was no longer a fixed entity, but rather a fluid and changing goal. Market research became focused on competitors' products and on what exactly customers meant when they described product quality, and quality was defined by the

life-cycle of a product and maintenance costs rather than its initial purchase price (Garvin, 1988).

As businesses began to pay attention to competitors' prices, so too did the competitor, setting off a cycle of continually raising the bar for who had the highest quality. Rather than shooting for acceptable levels of quality, the organization was now focused on a broader cycle of continuous improvement, which would necessitate the active involvement of executive management. Using a continuous improvement approach, "… required a dedication to the improvement *process* as well as the commitment to the entire company" (Garvin, 1988, p. 27).

One of the first forms of the continuous improvement approach developed from a book by William Ouchi in 1981. He wrote *Theory Z: How American Business Can Meet the Japanese Challenge.* Ouchi introduced industry to the concept of quality circles. In addition to industry, the U.S. Navy began to use quality circles to such a degree that they refined the process into a method they called *Total Quality Management*. This term has evolved into an umbrella term that covers many forms of quality management and continuous improvement in both private and public organizations. As the use of quality has evolved, so too has the terminology. Today, the term Continuous Quality Improvement is being used more often the Total Quality Management in matters of quality management (Ouchi [1981], as cited in Marchese, 1991).

As quality has become more strategic, linking more closely to profitability, business objectives, and the consumers' needs through continuous improvement, it

has spread to all levels of the organization. The Malcolm Baldrige National Quality Award reflects this new evolution.

Malcolm Baldrige National Quality Award (MBNQA)

History

As a result of the declining productivity in the world market in 1982, President Reagan signed legislation mandating a national study be conducted to determine the nation's ability to increase productivity and sustain itself against foreign competition. This legislation precipitated the American Productivity and Quality Center holding several computer networking conferences in preparation for a White House conference on productivity in 1983 (DeCarlo & Sterrett, 1990).

During this same timeframe, several other efforts were underway geared toward a similar purpose. The ASQC, with the help of Alvin Genneson, corporate vice president of quality at Revlon, was working on forming the National Advisory Council for Quality (NACQ) in an effort to develop a national awareness of quality and provide an advisory body on quality and productivity issues for all levels of industry and government (DeCarlo & Sterett, 1990). The goal of the NACQ, according to DeCarlo and Sterett, was, "… to become the recognized center for training, publications, conferences, and research in the quality disciplines," (p. 21). This body was officially formed in February of 1982.

While the corporate sector was working on their response to the productivity crisis, the American Productivity and Quality Center (APQC) was preparing for the upcoming White House Conference on Productivity. They held several computer

networking conferences between April and September of 1983. Approximately 175 leaders in quality attended the conferences (DeCarlo & Sterett, 1990). Some of the feedback from these leaders included recommendations for a national quality award much like the Deming Award in Japan and a national committee to coordinate the award, much like the Union of Japanese Scientists and Engineers. At the same time, the U.S. government was conducting a study and coming up with a similar suggestion. Table 1 is a timeline of coinciding events illustrating that the APQC was a few months ahead of the government in identifying and suggesting a solution to the national quality and productivity problem (DeCarlo & Sterett, 1990; Evans & Lindsay, 1999).

TABLE 1. **Political Timeline for the Creation of the Malcolm Baldrige National Quality Award**

Date	Event
February 1981	American Society for Quality Control and Alvin Gunneson led an effort that resulted in the creation of the National Advisory Council for Quality (NACQ).
October 1982	President Reagan signs legislation mandating a national study be conducted on productivity.
April–September 1983	American Productivity and Quality Center (APQC) conducted several computer networking conferences in which about 175 business executives and academicians came up with the idea that a national quality award, similar to the Deming Prize in Japan, was needed in the U.S.
September 1983	White House Conference on Productivity was held.

TABLE 1. Continued

Date	Event
December 1983	The National Productivity Advisory Committee (NPAC), appointed by the president, made the recommendation for the creation of a national medal for quality. The recommendation was tabled, however, due to a lack of direction for funding, format, criteria and other necessary details for administration.
April 1984	White House Conference on Productivity Report was published, issuing a challenge for improvement of productivity and calling for an annual national medal for productivity achievement.
September 1985	Formation of the Committee to Establish a National Quality Award comprised of private and academic sector members from ASQC, APQC, NASA, Ford Motor Co., and others (DeCarlo & Stennett, 1990)
June 1986	John Hudiburg, John Hansel, and Joseph Juran testified before congress on the potential for a national quality award
June 8, 1987	House of Representatives passed the National Quality Improvement Act of 1987
August 20, 1987	President Reagan signed the Malcolm Baldrige National Quality Improvement Act of 1987 into law.

Over the next few years, the quality productivity concern had become well-known and several groups were working together to come up with suggestions and solutions. Many of those suggestions focused on the creation of a national award. By September of 1985, the Committee to Establish a National Quality Award (later called the National Organization for the United States Quality Award) was created and a

focused effort on the quality award was underway. A year later, an initial draft of the criteria had been developed and support was growing from the President's office.

During this same time, an effort was being made to put legislative action in place for a national quality award. Florida Power and Light's (FPL) chairman and CEO, John Hudiburg and Marshall McDonald were key contributors to this effort. They met with Congressman Don Fuqua who was the chairman of the House Committee on Science and Technology at the time. In March of 1986, Fuqua was sending members of his staff to FPL and Japan to learn more about quality improvement and by June of 1986 the idea for a national quality award was formally discussed in a legislative committee meeting in Washington, DC (DeCarlo & Sterett, 1990).

After much political push and pull, a bill creating the National Quality Award was drafted and passed through the House in June of 1987. Shortly after it passed, a tragic accident occurred that took the life of Commerce Secretary Malcolm Baldrige. When the bill reached the Senate, they renamed the bill after Baldrige and passed it. The Malcolm Baldrige National Quality Improvement Act was signed into law as Public Law 100-107 by President Reagan on August 20, 1987.

The program would focus on the following points:

- Helping to stimulate American companies that improve the quality of and productivity for the pride of recognition while obtaining a competitive edge through increased profits
- Recognizing the achievements of those companies that improve the quality of their goods and services and providing an example to others
- Establishing guidelines and criteria that can be used by business, industrial, governmental, and other enterprises in evaluating their own quality improvement efforts
- Providing specific guidance for other American enterprises that wish to learn how to manage for high quality by making available detailed

> information on how winning enterprises were able to change their cultures and achieve eminence. (Evans & Lindsay, 1999, p. 115)

Administration

The MBNQA is housed under the National Institute for Standards and Technology (NIST) in the U.S. Department of Commerce. The mission of NIST is, "… to promote U.S. economic growth by working with industry to develop and apply technology, measurements, and standards" (NIST, 1998, p. 1). The MBNQA is one of the ways in which the NIST fulfills its mission. It is a public-private partnership, funded through a private foundation (Evans & Lindsay, 1999).

Criteria

Examination for the award is based on a set of criteria called the "Criteria for Performance Excellence." Juran (1997) stated that he believed as of the early 1990s, this criteria was the "… most complete available definition of TQM…" available to companies who were interested in introducing a total quality management system into their organization.

As the MBNQA was established, Curt Reinman, Award Program Director at NIST, began the development effort for the criteria. In 1988, he had resources such as the criteria from the Deming Prize in Japan, the NASA criteria and several others to draw upon. He also spoke with more than 70 quality experts. The result was the seven categories we know today, along with several sub categories and examination items (DeCarlo & Sterett, 1990). In the veign of continuous improvement, the sub-categories and items evolve each year based on feedback from the various applicants,

examiners, and judges. The seven main categories continue, however, to serve as the backbone of the MBNQA criteria. Reinmann pointed out, in an interview, that those at the NIST must keep an institutional mindset because they are dealing with such a large and diverse community. The categories are meant to be broad while the sub-categories and items were meant to evolve with the changing times (Bell & Keys, 1998).

Today, the Criteria for Performance Excellence are a set of expectations that define the critical success factors that drive organizational success. They are made up of 100 questions divided into an Organizational Profile and seven Categories by which examiners assess the organization's level of approach and deployment and results achieved (NIST, 1998). The Organizational Profile is a synopsis of the key influences on the organization and the key issues the organization faces. The language of the criteria and categories vary slightly by the different sectors. For the purpose of this review, the Business criteria and language will be used. The seven categories for Business are: Leadership, Strategic Planning, Customer and Market Focus, Information Analysis, Human Resource Focus, Process Management and Business Results (NIST, 1998).

Impact

Juran tells us that the MBNQA was a great stimulus for improving quality and spreading awareness (Juran, 1997). Although the annual number of applicants for the award is usually fewer than 100 (Juran, 1997), the NIST distributed over 800,000 copies of the criteria in 2001 (NIST, 1998). Clearly, organizations are realizing the

profound impact the Criteria for Performance Excellence can have when used as a guide for quality improvement efforts.

The establishment and success of the MBNQA has also led to the proliferation of other quality award programs. Numerous European and Latin-American countries and others have created national awards patterned after the Baldrige (Juran, 1997). There has also been a tremendous growth of several state quality awards. Forty-two states have quality award programs that use the Baldrige as a model (Network for Excellenc*e*, n.d.). Texas is one of those states. This study was designed to focus exclusively on the Quality Texas organization and its Texas Award for Performance Excellence. The remainder of this review will be a description of the assessment process using the Texas Award for Performance Excellence guidelines, which are extremely similar to the MBNQA assessment guidelines.

Texas Award for Performance Excellence

The Texas Award for Performance Excellence (TAPE) is awarded by the Quality Texas non-profit corporation known as the Quality Texas Foundation. The development of a quality award for the state was initiated by the Texas Governor's office in 1990 (Quality Texas Foundation, 2005). According to the Quality Texas Web site section on "About Quality Texas" (Quality Texas Foundation, 2005), cooperative efforts between the governor's office, the Texas Department of Commerce and other Texas businesses produced awareness seminars that were presented around the state to several hundred organizations. While awareness was growing, EDS Corporation was busy developing a state quality award with the input of leaders from state

government, business and education. The committee produced the TAPE and opened it to "… government, education, nonprofit, and business organizations" (Quality Texas Foundation, 2005). In 1994, Quality Texas was established as an independent corporation that would serve as the administrator of the award. The Quality Texas Foundation is now a recognized 501c3 nonprofit organization and headquartered in Dallas, Texas, (Quality Texas Foundation, 2005).

The Quality Texas Foundation defines quality as, "… the essential character of excellence and superiority. Quality is measured by all our customers and is critical for the success of any product, organization, or service company" (Texas Quality Foundation, 2005, p. 1).

The goal of the Quality Texas Foundation is to "… establish a greater awareness of quality principles in Texas," (Quality Texas Foundation, 2005, p. 1).

Application Process

The information on the TAPE application, scoring, and examiner selection and training found in the following paragraphs is taken from the Quality Texas Web site as of February 2005. This information reflects the most up-to-date language and most recent versions of the TAPE criteria except where reference is made to previous versions for TAPE materials or other resources.

Like the Baldrige Award, the TAPE process begins with an organizational self-assessment followed by a third-party examiner assessment and, in some cases, a site visit by the examiners. The TAPE varies from the MBNQA in that it is open to more sectors including manufacturing, service, small business/organization, public sector,

education, non-profit and healthcare (Quality Texas Foundation, 2005). Unlike the MBNQA which limits the number of awards granted to 3 per sector, there is no limit to the number of Texas awards. Additionally, the TAPE offers 3 options in the application process while the MBNQA offers only an award level application (Quality Texas Foundation, 2005).

There are three levels of application from which organizations can choose based on their level of experience with quality including Self-Assessment Level, Progress Level and Award Level. The highest level, which is the focus of this study, is the Award Level Process. The entire set of performance criteria used for the Baldrige Award is utilized in the Award Level. This level is mainly used by those organizations that have been using the principles and practices of performance excellence for a significant period of time (Quality Texas Foundation, 2005). Organizations applying at the level of award must begin by submitting an application (50-page limit) and an organizational profile (previously described in the section on the MBNQA) which gives examiners an overview of the particular organization, the issues it faces and its key business objectives. The application and organizational profile are made up of approximately 100 questions that, when addressed, help the applicant to thoroughly describe every aspect of the organization. The questions are written is such a way as to make them applicable to most organizations.

There are some cases in which different language is required due to the type of organization applying. For that reason, there are three different versions of the application including Generic, written for any for-profit organization; Education, written for all public or non-profit educational institutions; and Health Care, written

for agencies who deliver health care services directly to individuals. Although there are different versions of the language of the applications, the fundamental values and criteria measures are the same.

Each version of the TAPE application has core values written for that type of institution. Table 2 is a reflection that the core values, while worded slightly differently, are fundamentally the same.

TABLE 2. Core Values of the Texas Award for Performance Excellence

Generic	Education	Health Care
Visionary Leadership	Visionary Leadership	Visionary Leadership
Customer-Driven Excellence	Learning-Centered Education	Patient-Focused Excellence
Organizational and Personal Learning	Organizational and Personal Learning	Organizational and Personal Learning
Valuing Employees and Partners	Valuing faculty, staff, and partners	Valuing Staff and Partners
Agility	Agility	Agility
Focus on the Future	Focus on the Future	Focus on the Future
Managing for Innovation	Managing for Innovation	Managing for Innovation

TABLE 2. Continued

Generic	Education	Health Care
Management by Fact	Management by Fact	Management by Fact
Social Responsibility	Social Responsibility	Social Responsibility and Community Health
Focus on Results and Creating Value	Focus on Results and Creating Value	Focus on Results and Creating Value
Systems Perspective	Systems Perspective	Systems Perspective

The criteria are divided into the seven categories of the MBNQA. The categories for each type of institution are shown in Table 3.

TABLE 3. Categories for the Texas Award for Performance Excellence

Generic	Education	Health Care
Leadership	Leadership	Leadership
Strategic Planning	Strategic Planning	Strategic Planning
Customer and Market Focus	Student, Stakeholder, and Market Focus	Focus on Patients, Other Customers, and Markets
Measurement, Analysis, and Knowledge Management	Measurement, Analysis, and Knowledge Management	Measurement, Analysis, and Knowledge Management
Human Resource Focus	Faculty and Staff Focus	Staff Focus

TABLE 3. Continued

Generic	Education	Health Care
Process Management	Process Management	Process Management
Business Results	Organizational Performance Results	Organizational Performance Results

Each category consists of questions that are broken down into more specific "Items" and "Areas to Address" where more in-depth questions are answered. In 2005, the number of Criteria Items increased from 18 to 19 and the number of Areas to Address increased from 29 to 32 (Quality Texas Foundation, 2005). These changes reflect the efforts of the Quality Texas Foundation to continue to evolve with the changing needs and focus of organizations today. The 2005 Criteria changes address today's increased focus on governance and ethics, the need to capitalize on knowledge assets, the need to create value for customers and the business, and the alignment of all aspects of the performance management system (Quality Texas Foundation, 2005).

Once the application is submitted, a team of examiners assesses the organization based solely on what is written in the application and profile. The first step in the examination is to individually examine the documents and fill out a comprehensive scoring booklet. Next, the examiners meet as a team for the purpose of coming to consensus on every item of the scoring criteria. This is done in a consensus meeting that may take place via phone conference or over a weekend. Consensus meetings often take up to eight hours as examiners must discuss and come to consensus on

each Item and Area to Address and overall Category Score and identifies site visit issues. In this way, the examination process taps the expertise and experience of all of its examiners and comes to as objective of an assessment as possible.

Once the examiner team comes to consensus, a Panel of Judges determines which applicants will be awarded site visits. Those applicants who do not receive site visits will receive a detailed feedback report which outlines organizational strengths and opportunities for improvement. Those applicants who receive a site visit, which is simply a visit to the organization for the purpose of verifying the application, will receive a feedback report after the visit is completed and the examiner team has had more time to review their findings. The feedback report is an invaluable resource for each organization as a continuous improvement tool.

A final report of all evaluations of the applicants receiving site visits is conducted by the Panel of Judges who develop recommendations and pass them to the Board of Directors. The Board of Directors who, along with the Program Manager, determine who will receive the TAPE (Quality Texas Foundation, 2005).

Scoring Process

As an examiner reads an application and evaluates the application based on the item questions, he/she writes comments, keeping track of the strengths and opportunities for improvement that he/she sees in that section. The examiner will also write summative comments on key factors found in the category which represent a broader overall picture of what he or she sees in that category. All these comments are used by the examiner to help in deciding what score to give a particular item and

category. Tables 4 and 5 are the "Scoring Guidelines Quick Cards" for evaluating Approach and Deployment items and Results items. This Quick Card helps the examiner decide what score to give an item. Scores can only be given in increments of ten on a scale of 1-100. If, therefore, an examiner feels that an organization falls in the 30% - 40% range, he/she must pick either a score of 30 or 40 to reflect that the organization falls in either the low or high end of the 30% - 40% range.

TABLE 4. TAPE Scoring Guidelines for Approach/Deployment Items (for Use with Category 1-6 Items)

Score	Approach			Deployment
	Appropriateness to Requirements	Effective & systematic	Alignment	
0%	Information is anecdotal.	No systematic approach is evident.	--	--
10% to 20%	Responsive to the basic requirements of the item.	The beginning of a systematic approach is evident. Early stages of transition from reacting to problems to a general improvement orientation are evident.	--	Major gaps in deployment that would inhibit progress in achieving the basic requirements of the item.

TABLE 4. Continued

Score	Approach			Deployment
	Appropriateness to Requirements	**Effective & systematic**	**Alignment**	
30% to 40%	Responsive to the basic requirements of the item.	An effective, systematic approach is evident. The beginning of a systematic evaluation and improvement is evident.	--	The approach is deployed, although some areas or work units are in early stages of deployment.
50% to 60%	Responsive to the overall requirements of the item and key business requirements.	An effective, systematic approach is evident. A fact-based, systematic evaluation and improvement system is in place for improving the efficiency and effectiveness of key processes.	The approach is aligned with basic organizational needs identified in other Criteria Categories.	The approach is well deployed, although deployment may vary in some areas or work units.
70% to 80%	Responsive to the multiple requirements of the item and to current and changing business needs	An effective systematic approach is evident. A fact-based systematic evaluation and improvement process and organizational learning/sharing are key management tools; there is clear evidence of refinement, innovation, and improved integration as a result of organizational-level analysis and sharing.	The approach is well integrated with organizational needs identified in other Criteria Categories.	The approach is well-deployed, with no significant gaps.

TABLE 4. Continued

Score	Approach			Deployment
	Appropriateness to Requirements	**Effective & systematic**	**Alignment**	
90% to 100%	Fully responsive to all requirements of the item and all current and changing business needs.	An effective, systematic approach is evident. A very strong fact-based, systematic evaluation and improvement process and extensive organizational learning/sharing are key management tools; strong refinement, innovation, and integration, backed by excellent organizational-level analysis, are evident	The approach is fully integrated with organizational needs identified in the other Criteria Categories.	The approach is fully deployed without significant weaknesses or gaps in any areas or work units.

TABLE 5. Scoring Guidelines for Results Items (for Use with Category 7 Items)

Score	Current Performance	Trends	Comparisons	Breadth & Importance
0%	There are no results or poor results in areas reported.	--	--	--
10% to 20%	There are some improvements and/or early good performance levels in a few areas.	--	--	Results are not reported for many to most areas of importance to the organi-zation's key business requirements.

TABLE 5. Continued

Score	Current Performance	Trends	Comparisons	Breadth & Importance
30% to 40%	Improvements and/or good performance levels are reported in many areas.	Early stages of developing trends.	Early stages of obtaining comparative information.	Results are reported for many to most areas of importance to the organization's key business requirements.
50% to 60%	Improvement trends and/or good performance levels are reported for most areas.	No pattern of adverse trends and no poor performance levels are evident in areas of importance to the organization'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.	Results reported for most areas of importance to organization's key requirements. Results address most key customer, market, and process requirements.
70% to 80%	Current performance is good to excellent in areas of importance to the organization's key business requirements.	Most improvement trends and/or current 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/or very good relative performance levels.	Results reported for most areas of importance to your organization's key requirements. Results address most key customer, market, process, and action plan requirements.

TABLE 5. Continued

Score	Current Performance	Trends	Comparisons	Breadth & Importance
90% to 100%	Current performance is excellent in most areas of importance to the organization's key requirements.	Excellent improvement trends and/or sustained excellent performance levels are reported in most areas.	Evidence of industry and benchmark leadership is demonstrated in many areas.	Results reported for most areas of importance to your organization's key requirements. Results fully address key customer, market, process, and action plan requirements.

Once examiners have scored individual items on the 10-point scale and have checked to see that their scores are in alignment with their comments on strengths and opportunities for improvement, they fill out the score sheet by transferring their item scores and figure the percentages eventually arriving at a score for each category and a total score for the organization. Table 6 is a reflection of the total possible point value for each item and category for a total possible point score of 1000.

TABLE 6. Award Level Criteria for Performance Excellence – Item Listing

Categories/Items	Point Values
P Preface: Organizational Profile	
P.1 Organizational Description	
P.2 Organizational Challenges	

TABLE 6. Continued

Categories/Items		Point Values
1 Leadership		**120**
1.1 Organizational Leadership		
1.2 Social Responsibility	70	
	50	
2 Strategic Planning		**85**
2.1 Strategy Development	40	
2.2 Strategy Deployment	45	
3 Customer and Market Focus		**85**
3.1 Customer and Market Knowledge	40	
3.2 Customer Relationships and Satisfaction	45	
4 Measurement, Analysis, and Knowledge Management		**90**
4.1 Measurement and Analysis of Organizational Performance	45	
4.2 Information and Knowledge Management		
	45	
5 Human Resource Focus		**85**
5.1 Work Systems	35	
5.2 Employee Learning and Motivation	25	
5.3 Employee Well-Being and Satisfaction	25	
6 Process Management		**85**
6.1 Value Creation Processes	50	
6.2 Support Processes	35	
7 Results		**450**
7.1 Customer-Focused Results	75	
7.2 Product and Service Results	75	
7.3 Financial and Market Results	75	
7.4 Human Resource Results	75	

TABLE 6. Continued

Categories/Items	Point Values
7.5 Organizational Effectiveness Results	75
7.6 Governance and Social Responsibility Results	75
TOTAL POINTS	**1000**

Table 7 is a reflection of the Score Summary Worksheet, wherein examiners transfer item and category point scores and multiply by a percentage to get a total percent score. These scores are sent to the team leader who compiles all examiner scores and keeps them for the consensus meeting.

TABLE 7. Score Summary Worksheet—Generic Criteria

Summary of Criteria Items	Total Points Possible A	Percent Score 0-100% (Stage 1—Use 10% Units) B	Score (A x B) C
Category 1			
Item 1.1	70	10%	
Item 1.2	50	10%	
Category 1 Total	**120**		
			SUM C
Category 2			
Item 2.1	40	%	
Item 2.2	45	%	
Category 2 Total	**85**		
			SUM C

TABLE 7. Continued

Summary of Criteria Items	Total Points Possible A	Percent Score 0-100% (Stage 1—Use 10% Units) B	Score (A x B) C
Category 3			
Item 3.1	40	%	
Item 3.2	45	%	
Category 3 Total	**85**		
			SUM C
Category 4			
Item 4.1	45	%	
Item 4.2	45	%	
Category 4 Total	**90**		
			SUM C
Category 5			
Item 5.1	35	%	
Item 5.2	25	%	
Item 5.3	25	%	
Category 5 Total	85		
			SUM C

TABLE 7. Continued

Summary of Criteria Items	Total Points Possible A	Percent Score 0-100% (Stage 1—Use 10% Units) B	Score (A x B) C
Category 6			
Item 6.1	50	%	
Item 6.2	35	%	
Category 6 Total	**85**		
			SUM C
Category 7			
Item 7.1	75	%	
Item 7.2	75	%	
Item 7.3	75	%	
Item 7.4	75	%	
Item 7.5	75	%	
Item 7.6	75	%	
Category 7 Total	**450**		
GRAND TOTAL (D)	**1000**		

Examiner Selection and Training

This study was designed to analyze the TAPE process at the level of examiners. With both the TAPE and MBNQA, this volunteer group is known as the Board of Examiners. The Board is made up of experienced professionals from the private, public, education, and health care sectors and is selected by the Selection Team made up of members of the Board of Overseers, Judges, and staff members from Quality Texas. Examiners are chosen based on their personal qualifications from past and

present work experience. They cannot be considered affiliates of their organizations or representatives of their employers.

Examiners must possess certain general qualifications including the following:

- Broad knowledge of quality and performance excellence principles
- Length, breadth, and type of experience
- Analytical skills
- Communication skills
- Education and training
- Achievements and recognition
- Ability to meet rigorous time commitments as scheduled, or when called upon

The Board of Examiners consists of approximately 150 members and is responsible for evaluating applications, preparing feedback reports, conducting site visits, and making recommendations to the Board of Directors (Quality Texas Foundation, 2005). In addition to examining an organization, there are several different roles and responsibilities that members may have including serving as a Team Leader or Feedback Writer.

Examiner teams are a heterogeneous mix of new, returning, and senior examiners made up of men and women of various ages and experience levels. New examiners are working with the TAPE for the first time, returning examiners have one or two years of experience, and senior examiners are considered veterans not only in the TAPE process, but also in the field of quality. All examiners, regardless of experience, must attend a rigorous three-day examiner training each year. Prior to

attending training, examiners must conduct a full examination of a faux organization to be used as a case study during the three days of training. During the training, examiners work in teams to analyze the case study and learn about various items within the categories of the TAPE, how to write non-prescriptive feedback, and how to identify strengths and opportunities for improvement regarding various criteria items. They also learn how to score items and come to consensus with other examiners on their team. This process of using outside examiners to assess organizations is known as "third-party assessment" and is the focus of this study.

Examiner Training, Accuracy, and Reliability

This portion of the review of literature will address the research on rater training and how it affects accuracy and reliability of individuals and groups. For this section, the terms "rater" and "examiner" will be used interchangeably.

Research on third-party assessment and accuracy as well as organizational assessments used as evaluation tools is relatively non-existent. One of the few sources of discussion on the lack of research comes from a forum held by the European Organization for Quality and can be found in the Proceedings of the First European Forum on Quality Self-Assessment held in 1994 in Milano, Italy. The theme of this forum was: The Use of Quality Award Criteria and Models for Self-Assessment Purposes. Several contributors at this forum expressed concern over the assumption that criteria for quality awards were being used in self-assessments and resulted in business decisions that were being made without existing research to prove that this criteria, in fact, produced accurate and reliable results (Conti, 1994; Fuchs &

Stuntebeck, 1994; Jernberg et al., 1994; Martellani, 1994). Tito Conti (1994) highlighted this concern in the forward of the proceedings when he called for, "… a critical review of self-assessment criteria and methodologies" (p. 5). More recently, Evans and Jack (2003) conducted a study that attempted to validate some of the linkages between the Baldrige criteria and business outcomes and noted that little empirical research had been performed to validate the Baldrige criteria and its core concepts and values.

While examiner training research yields few results, however, performance appraisal rater training, found in the psychology literature, has experienced more scrutiny (Coleman et al., 2001). This researcher was able to find only one study conducted specifically on training and scoring organizational self-assessments where third-party examiners and quality award systems like the MBNQA were used in the research. Garry Coleman et al. conducted a study in 2001 which they say is the, "… first known application of accuracy indices to the scoring of organizational assessments," (Coleman et al., 2001, p. 523).

Research found in the industrial organizational psychology literature refers to "rater" accuracy in the context of rating human performance and the training of those raters. Although this research does not directly relate to external examiner groups evaluating an organization of which they are not members, it does indeed indirectly relate to the third-party examiner accuracy and interrater reliability.

It is surprising to find such little inquiry in the area of organizational assessment examiner training. With the exception of the concerns expressed about accuracy of examiner scores published in the papers of the First European Forum on Quality Self-

Assessment (Conti, 1994; Fuchs & Stuntebeck, 1994; Jernberg et al., 1994; Martellani, 1994), little else has been written. Quality awards based on organizational self-assessments are growing in popularity all over the world and are seen as highly reliable sources of information regarding successful business practice. Some of the top companies in the world use feedback from the MBNQA and other quality awards in their efforts to improve, recognize management practices, and make decisions, yet few have ever questioned the feedback or results of the examination process. Coleman et al. (2001) believe it is important to establish reliability and validity of the examination process in order to gain credibility and continue to improve the system.

The methods used by quality award organizations to address the issue of potential errors in examiner scores is to require extensive training for the examiners and to create heterogeneous teams of experts in the area of quality through the use of selection criteria (Godfrey & Meyers, 1994; NIST, 1998). While the length and intensity of training for quality programs varies, the focus of this study is the Texas Award for Performance Excellence (TAPE), which follows the MBNQA case study and three-day training method.

Scoring for the TAPE begins with an applicant organization completing a self-assessment. Examiners then review the self-assessment document and assign scores to the organization in the seven categories described previously in this review. A major element of the examination is the feedback provided by the examiners to the organization. It is this feedback that the organization uses for decision-making and improvement efforts. Consequently, it is important that feedback be accurate and reliable.

As previously stated, much study has been conducted on performance appraisal rater training. The two most common forms of rater training are Frame of Reference (FOR) training and Rater Error Training (RET) (Coleman et al., 2001; McIntyre, Smith & Hasset, 1984; Stamoulis & Hauenstein, 1993). FOR training gives raters a common "frame of reference" so that a variety of raters can evaluate the same worker behaviors and come to similar conclusions (McIntyre et al., 1984). Training includes examples of job performance being shared with raters along with the "true" ratings that should be assigned to the example performance. The "true" rating is based on what expert raters have concluded to be the appropriate score. FOR is the type of rater training that most closely approximates examiner training for the MBNQA. Additionally, FOR's approach of using expert raters to access a "true score" is the basis for using the consensus score as the "true score" in this researcher's study.

RET is the method in which raters are provided training on common errors such as being overly lenient or severe, the halo effect, central tendency, and contrast errors (Smith, 1986). Once made aware, raters are "admonished" to avoid these psychometric errors (McIntyre et al., 1984). According to McIntyre et al. (1984) and Stamoulis and Hauenstein (1993), RET tends to reduce error in the assessment of individual performance.

An additional type of rater training which has bearing on this research is Performance Dimensions Training (PDimT) (Coleman et al., 2001). PDimT studies, "… attempt to improve the effectiveness of ratings by familiarizing raters with the dimensions by which the performance is rated. This is done by providing descriptions

of job qualifications, reviewing the rating scale used in the evaluations, or having raters participate in the actual development of the rating scale," (Smith, 1986, p. 30).

According to Coleman et al. (2001), "Rater training generally improves one or more aspects of rater effectiveness, but may result in degradation or no change to other aspects" (p. 516). Smith suggests that rater outcomes are improved when raters are given opportunities to be more actively involved in the rating process (Smith, 1986). Evidence from Smith's study further suggests that combining two training approaches, such as FOR and PDimT, will increase accuracy (Coleman et al., 2001; Smith, 1986). Coleman et al. (2001) point out that quality award examiners usually receive FOR training and little PdimT. According to Coleman et al. (2001, p. 517), "There appears to be an implicit assumption that those selected as evaluators already have knowledge of the performance dimensions and do not require PdimT."

All forms of training have strengths and weaknesses regarding scoring accuracy. As observed by Coleman et al. (2001), "Accuracy may be viewed as the relative absence of error, where error is deviation from the true scores of organizational performance. Accuracy of scores can be measured by examining the relative distance between an evaluator's scores and the true scores of organizational performance" (p. 514).

There are two main descriptors of scoring accuracy when talking about organizational evaluation; elevation and dimensional accuracy (Coleman et al., 2001; Hauenstein & Alexander, 1991). Elevation (EL) is the difference between the average of scores for an examiner and the average of the true scores for a given set of criteria (Coleman et al., 2001). Dimensional Accuracy (DA) "… measures the accuracy with

which an evaluator scored a single organization on a set of related dimensions," (Coleman et al., 2001, p. 515). According to Hauenstein and Alexander (1991), a perfect scenario for EL accuracy would be reached when a rater's average observed rating equaled the average of the target scores. Coleman et al. (2001) point out that it is sufficient, but not necessary, to have a correlation of positive one between a rater's observed ratings and the target ratings, as well as a rater's variance to equal the variance of the target scores for an ideal DA score.

Different types of training are more or less suitable depending on the intent of the organizational assessment. Stamoulis and Hauenstein (1993) noted that FOR training was better for increasing dimensional accuracy and RET was better for elevation. They cited Murphy, Garcia, Kerkar, Martin, and Balzer (1982) in noting the importance of considering what type of organizational decisions are being made as that will, or at least should, have an impact on which type of accuracy to emphasize (Stamoulis & Hauenstein, 1993). Coleman et al. (2001) state that EL is best used when examiner scores are being used to ascertain whether an organization meets a particular level of performance. They further submit that because an EL score is an indicator of how close an examiner's score is to the true score, it would be useful in deciding whether examiners' scores are accurate enough to be used for decision making. Additionally, Coleman et al. state that DA is useful in the feedback process when examiners are identifying strengths and weaknesses of an organization. Therefore, knowing the DA for a given set of examiners will tell whether the identified strengths and weaknesses are accurate and reliable for useful decision making (Coleman et al., 2001). Given the findings about the strengths of FOR

training, one could assume that quality award or third-party examiners tend to score organizations with better dimensional accuracy. This makes sense when recollecting the training process used by the MBNQA.

Stamoulis and Hauenstein (1993) suggest FOR training for the improvement of DA as well as other forms of training so that all areas of accuracy are addressed. Coleman et al. (2001) add to the discussion by suggesting, “careful selection of evaluators to improve the dimensional accuracy of organizational assessment scoring“ (p. 524). They go on to point out that the MBNQA has the “luxury” of selecting examiners from a large pool of quality experts making it possible to compensate for less training while other quality awards do not. Therefore, smaller quality award programs like the TAPE may need to pay more attention to the type and variety of training they provide.

Group Effect on Rater Accuracy

An important aspect of the quality award evaluation process is the consensus meeting and generation of the consensus scores. Once examiners have completed their individual assessments of the quality award applicant, they meet with their team to discuss each item score and category score. It is believed that a heterogeneous group of quality experts, having read the same organizational profile and application, will each see different aspects and dimensions of the organization’s performance across the seven categories and nineteen items. As they discuss what they see as the strengths and weaknesses of the organization’s approach, deployment and results, they will be more likely to gain a more accurate picture of the organization as a

whole, or will be able to generate deeper clarifying questions to be brought up at a site visit.

Research on the effect of organizational assessment accuracy for groups, again yields few results (Martell & Borg, 1993). This researcher, however, did find a study that focused on behavioral rating accuracy of groups. While generalizations should be made with caution, there does appear to be some important information to inform this study as it relies on the consensus scores of the TAPE examinations as the true score against which individual rater accuracy will be measured.

The notion that groups of raters will generate different performance ratings than individuals is not remarkable. Wherry and Bartlett (1982) suggested that multiple raters would demonstrate more accuracy and less bias than individual raters. Questions still exist, however, as to whether or not this is true. Few studies have been conducted to flesh out this question. There are both positive and negative aspects to group raters. The following section will highlight those aspects.

Assets of Group Raters

Martell and Borg (1993) suggest three areas where groups may be more accurate in rating performance. Performance assessment processes, as with organizational assessment processes, often contain delays between the time a rater observes and the time the rater submits an evaluation. Working with a group increases the probability that at least one member of the group will remember an important detail and will be able to discuss that with the group, thereby “refreshing” memories and increasing the potential for accuracy (Martell & Borg, 1993). The assumption here is that the rater

who remembers the detail remembers the behavior accurately and does not influence the group with opinions.

Group decisions are typically the result of a previous discussion where there is "give-and-take" (Martell & Borg, 1993, p. 43), which produce more critical thinking and commitment to the task. These discussions create a sort of accountability, which according to Martell and Borg (1993), results in information being processed more carefully. Additionally, lengthy discussion promotes better memory accuracy as members of the group spend more time searching their memories during the discussion.

Finally, individual-level errors stand a better chance of being corrected during the group discussion. Martell and Borg (1993) point out that correspondence bias and the consensus underutilization effect have been corrected through group discussion.

Liabilities of Group Raters

While raters working together in groups have great potential to remember more and decipher more ambiguity, they are also susceptible to other types of error. For example, groups have the potential to amplify biases of individuals. Martell and Borg (1993) suggest that groups are "… more susceptible to the representativeness heuristic insofar as they exhibit even greater reliance on individuating information (and less on base-rate information) than do individuals" (p. 43).

Tindale (1989) conducted a study on group raters and found that there is the potential to exaggerate the decision criteria adopted by individuals. He found that, when encouraged to adopt a particular bias, the group actually amplified individual-

bias when the situation provided no feedback for correction. The group outcome resulted in more errors in evaluation than individuals. Tindale subsequently suggested that "…any advantages of using groups for personnel … decisions may be limited to conditions where outcome feedback is … available" (p. 468). The fact that groups may amplify individual-level biases implies that groups may, in fact, be more biased than individuals (Martell & Borg, 1993).

Martell and Borg's (1993) research showed that in situations where there was a delay between the time of observation and the assessment, groups were able to remember behaviors more accurately. Groups also demonstrated, however, a greater bias than did individual raters (Martell & Borg, 1993). They suggest that this bias is similar to the "polarization effect" found in attribute research. They describe this effect with the following, "… during discussion, group members hear an increased number of arguments that favor the initial predisposition of most group members; consequently, there is a marked shift (polarization) in this direction," (1993, p. 47). Martell and Borg also suggest Social Comparison Theory as a possible explanation of group bias. Social Comparison Theory says that group members may experience public pressure to conform to the prevailing opinion of the group (Martell & Borg, 1993).

Research Methodologies Similar to the TAPE process

Delphi Method

While the validity of the TAPE process would be strengthened through further research, it is important to point out the similarities this process has to established

research methodologies that are already widely accepted as consistent and reliable and scientific. Two of these methodologies are the Delphi Technique and general qualitative research methods.

The Delphi is, "… a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (Linstone & Turoff, 1975, p. 3). It was developed in the 1950s as a result of a RAND Corporation project sponsored by the United States Air Force which, like the TAPE scoring process, was designed to establish a consensus of expert opinions through questionnaires and controlled feedback (Linstone & Turoff, 1975). The justifications for using such an approach resided in the fact that a more traditional method using data-collection and computer models was too extensive for the capabilities of the era, and even with a traditional approach, there were too many subjective factors that could not be eliminated. Today, there are still instances that require a Delphi approach. Researchers continue to encounter situations where accurate information is difficult or impossible to obtain and where subjectivity is an influencing factor in the research process (Linstone & Turoff, 1975).

The TAPE process of bringing examiners to consensus differs from the Delphi technique in the way that they establish group consensus. The Delphi method typically involves either a paper-and-pencil approach or a real-time approach of written responses through the use of computers. The researcher facilitates the summary of responses and redistribution of the results. This method uses written answers so that no one person can dominate the group (Miller & Salkind, 2002). By

eliminating face-to-face interaction, the Delphi method attempts to eliminate the influences of interpersonal interactions (Miller & Salkind, 2002).

The TAPE process differs from the Delphi in that all examiners are required to meet, either in person or by telephone conference so that each may be involved in an open dialogue. All examiners discuss their opinions and perceptions of the applicant and work toward a group consensus. Their conversation is facilitated by the team leader who is also one of the examiners. Unlike the Delphi technique, the TAPE process leaves itself open to the possibility of interpersonal influence having an impact on the outcome of an organization's overall score. The TAPE attempts to control for the possibility influence by careful screening of examiners to make sure that no examiner has a conflict of interest and by dispersing examiners by experience, gender, and other factors to create diverse teams.

Although the procedures of the Delphi and TAPE process differ, the end goal and underlying assumption that agreement among experts is the best way to estimate the value of a vague or unfamiliar variable (Miller & Salkind, 2002) are remarkably similar. The similarities between the TAPE consensus procedure and the widely-accepted research method called the Delphi lends credibility to the TAPE's scoring process.

Qualitative Research

Qualitative research, although not as obviously similar, does share some common elements with the TAPE consensus process in deriving conclusions from subjective data. There are many "definitions" of qualitative research as it means different things

to different people. Generally, qualitative research refers to research that is not derived through statistical methods (Auerbach & Silverstein, 2003; Emory & Cooper, 1991; Straus & Corbin, 1998). Just as there are a number of different quantitative research methods, there are also a number of qualitative research methods. Not all of these methods can be found in the TAPE scoring process. There are certain themes, however, that are common throughout qualitative research and lend themselves to the underlying philosophy of the TAPE process as well.

According to Auerbach and Silverstein (2003), two themes that are part of the qualitative research paradigm include using research participants as expert informants and the explicit use of the researcher's subjectivity and values. Qualitative researchers recognize that those who have had real life experience with a certain phenomenon, can and should be recognized as experts and as such, have valuable insights to offer (Auerbach & Silverstein, 2003). This philosophy feeds the qualitative approach to building credibility. Again, there are many methods of attending to credibility in the qualitative research arena. However, two methods share a common thread with the TAPE process.

Triangulation involves, "The use of evidence from different sources…and of different investigators…" (Robson, 1993) to enhance credibility. The TAPE process uses the individual scores from different examiners to enhance credibility. In addition, TAPE examiners review organizational assessment documents and conduct interviews during site visits which gives them a broader perspective of the organization. Both methods see different perspectives of investigators as a way to strengthen inquiry. Through the TAPE process, examiners' ability to accurately

assess organizations is strengthened through a shared understanding of how the organization operates.

A second element which builds the credibility of a qualitative research study is known as peer debriefing. Robson (2003) defines peer debriefing as the act of, "Exposing one's analysis and conclusions to a colleague or other peer on a continuous basis…." This technique is remarkably similar to the consensus building phase of the TAPE scoring process where examiners debrief their own thoughts and perspectives to each other in an effort to come to a common and enhanced understanding of the applicant organization.

Conclusion

Understanding the potential assets and liabilities of group raters has important implications for this study. The consensus process for the TAPE relies upon the team of examiners to come to consensus over 19 items and 33 areas to address in an effort to see all the strengths and opportunities for improvement in an organization. Careful selection of examiners and thoughtful assembly of teams along with intensive three-day training sessions are clearly beneficial in reducing error and increasing both elevation and dimensional accuracy. The opportunity for strong-minded and highly experienced examiners, however, to influence other less experienced examiners is a real possibility.

Looking at the TAPE consensus building process through a Delphi or Qualitative Research lens brings a different perspective, and therefore an even greater certainty that more study needs to be conducted on the ability of third-party examiners to

accurately assess organizations. It is therefore, important to understand the variation between and among examiners in order to address the potential for error or bias and to improve examiner training. This study was an attempt to make those comparisons and, through analysis, offer suggestions for improved examiner training.

CHAPTER III

RESEARCH METHODOLOGY

Introduction

Research, defined at its most basic level, is a systematically conducted inquiry that provides information for the purpose of solving problems (Emory & Cooper, 1994). There are many ways to categorize research; pure or applied, complex or simple, empirical or descriptive (Emory & Cooper, 1991). Whether the research follows a classical pattern of *a priori* hypothesis and testing or is in an earlier stage of discovery where descriptive or exploratory research is necessary, one thing all scientific research should have in common is that it should provide an answer to some question (Creswell, 1994; Dubin, 1978; Emory & Cooper, 1991; Miller & Salkind, 2002). Quality research adheres to standards of scientific method. According to Emory and Cooper (1994), these standards include a clearly defined purpose, sufficiently described procedures that allow for replication, careful planning, honest reporting, sufficient and appropriate analysis, objective and accurate conclusions, and must be conducted in such a way as to promote confidence in the integrity of the study and the researcher. Quantitative methods were used in this exploratory study to estimate the scoring stability of third-party examiners through all assessments conducted by examiners for the Texas Award for Performance Excellence during the years 2001 through 2004. The process for determining the reliability of the data as well as the specific quantitative methods used to answer each research question is outlined in this chapter.

Purpose

This study was an analysis of the scoring process for the Texas Award for Performance Excellence (TAPE) from the point at which an organization was scored by a team of Examiners to the point where a team consensus score was calculated and the application was forwarded on to the Panel of Judges. The purpose of this study was to determine the scoring stability of third-party examiners who were assessing organizational performance for the Texas Award for Performance Excellence. Results were used to formulate the implications of the obtained stability for Examiner training. Included in this chapter is a description of the population, the scoring process and an explanation of the statistical method used for the analysis of each research question.

Research Questions

In order to fulfill the purpose of this study, the following research questions were addressed.

1. Is the mean of the deviations of individual total scores from team total consensus score equal to zero?
2. Is the mean of the deviations of individual item scores from team item consensus scores equal to zero?
3. Do item deviation scores vary across the following classifications:
    a. Levels of Examiner Experience
    b. Sectors

c. Levels of Self Assessment

d. Levels of Team Experience

Operational Definitions

The terms below were used in this research study based on the corresponding definitions. They are repeated here from Chapter I to aid in ease of referral.

Scoring Stability—refers to the consistency in item scores across several examiners for the Texas Award for Performance Excellence.

Levels of Team Experience—3 levels of team experience (Senior, Average, and New) were developed including teams with 51% or more senior examiners, teams with 51% or more new examiners, and teams which were 50% new examiners and 50% senior examiners and teams with 51% or more new examiners. Returning examiners were combined with senior examiners for this research study.

Texas Award for Performance Excellence (TAPE)—the non-profit organization in the State of Texas that assesses organizational performance based on the quality philosophy and seven categories used in the Malcolm Baldrige National Award for Quality (Quality Texas Foundation, 2005).

Malcolm Baldrige National Quality Award (MBNQA)—the award program governed by the National Institute for Science and Technology in the United States that assesses and recognizes organizational performance based on the approach, deployment, and business results of quality principles. The MBNQA is the leading model for quality awards around the world (NIST, 1998).

Third-Party Examiner—an individual who has completed the TAPE training and has read and assessed an organization's application to the TAPE process (Quality Texas Foundation, 2005).

Category—one of seven areas addressed on the organizational assessment. Categories for the Baldrige and TAPE assessment include Leadership, Strategic Planning, Customer and Market Focus, Information and Analysis, Human Resource Focus, Process Management, and Results (Quality Texas Foundation, 2005).

Embedded Item—sub-categories within a category

Sector—the differentiation between various types of organizations. Sector titles for the Baldrige and TAPE include Small Organizations, Manufacturing, Education, Health Care, Public Business, and Service.

Self-Assessment Score—the score an examiner assigns to him/herself regarding his/her level of confidence and ability to assess an organization

Individual Total Score (ITS)—the final score given by one examiner to an organization prior to the team consensus meeting.

Total Consensus Score (TCS)—the score arrived at through consensus of all examiners on a team who assessed a particular organization. For the purpose of this study, the total consensus score will be used as the "true score" against which individual total scores will be measured in regard to Research Question 1.

Individual Total Deviation Score (ITDS)—the score produced by subtracting the total consensus score from the individual total score. [ITDS = ITS – TCS]

Team Mean Deviation Score (TMDS)—the mean of the individual total deviation scores for a team.

$$TMDS = [(ITS\text{-}TCS_1) + (ITS\text{-}TCS_2) + (ITS\text{-}TCS_3)\ldots..+ (ITS\text{-}TCS_n)] / n$$

Individual Item Score (IIS)—the score given by one examiner for each of the 17 embedded items in an organizational assessment. This score is given prior to the team consensus meeting.

Team Item Consensus Score (TICS)—the score given to an item by the team of examiners as a result of a consensus meeting. For the purpose of this study, the total consensus score will be used as the "true score" against which individual item scores will be measured in regard to Research Question 2.

Item Deviation Score (IDS)—the score produced by subtracting the team item consensus score from an individual item score. (IDS = IIS – TICS)

Item Mean Deviation Scores (IMDS)—the mean of the item deviation scores for a team. There are 17 item mean deviation scores for each team.

$$IMDS = [(IIS\text{-}TICS_1) + (IIS\text{-}TICS_2) + (IIS\text{-}TICS_3)\ldots..+ (IIS\text{-}TICS_n)] / n$$

Scoring Process

Examiners rated the applicant organizations from six different sectors including service, health care, education, small organizations, public organizations, and manufacturing organizations who applied for the Texas Award for Performance Excellence between 2001 and 2004. All 34 applicants, regardless of sector, filled out an organizational profile and a self-assessment document based on the TAPE/MBNQA Criteria and was limited to 50 pages.

In conducting the scoring process for the TAPE, examiners were divided into teams consisting of 7 to 10 members. Each team was assigned to a different

organization. The Quality Texas Foundation has a process for assigning examiners to teams so as to make the teams as balanced and non-biased as possible. The process begins with a pool of volunteer examiners who have applied and been selected to serve for the given year. Using the North American Industrial Classification System Code (NAICSC), a report is generated that matches examiners (based on experience) to organizations. Next, team leaders and feedback writers are chosen from senior and returning examiners who would be appropriate for examining the applicant organization. Remaining examiners are assigned to each team based on sector experience with an effort to balance senior, returning and new examiners and male or female examiners. Other examiners are chosen (for the purpose of establishing diversity within the team) based on alternative sector experience. The preferred number of examiners on a team is 7 to 8. When there are many more examiners than applicants, however, there may be more than eight examiners assigned to a case.

All examiners in this study rated organizations according to criteria in seven categories regardless of the sector in which the organization resided. The seven categories included Leadership (1.0), Strategic Planning (2.0), Customer and Market Focus (3.0), Information and Analysis (4.0), Human Resource Focus (5.0), Process Management (6.0), and Results (7.0). Within each category, there were 2 to 4 subcategories called “items.” Each examiner gave a score for each item within a category. Consequently, each examiner assigned 20 item scores to an organization before assigning the overall score.

Item scores were assigned in 10-point increments ranging from 10 to 100 and reflected an individual examiner’s assessment of the organization based on how the

applicant addressed the criteria in answering the item questions. For example, under the category of Leadership in 2004 there were two sub-questions, each with two items creating a total of four items to be addressed. Each item consisted of a group of questions that the applicant answered and the examiner used to assess the applicant's conformance to the item. The examiner, having studied the organizational profile and 50-page application, gave a score to the organization reflecting that examiner's assessment of the organization's level of quality management in the area addressed by the questions. For example, under the 2004 Leadership category (1.0), Item 1.a, the questions were, "How do senior leaders set and deploy organizational values, short- and longer-term directions, and performance expectations? How do senior leaders include a focus on creating and balancing value for customers and other stakeholders in their performance expectations? How do senior leaders communicate organizational values, directions, and expectations through your leadership system, to all employees, and to key suppliers and partners? How do senior leaders ensure two-way communication on these topics?"

Each examiner reviewed an organizational self-assessment, assigned points to each of the criteria items and then generated total score for the organization. The examiner then filled out the score sheet (see Table 6) and submitted it to the team leader to be used during the consensus meeting and for report writing.

Items were assigned points in increments of 10 on a 100-point scale. However, every item was not equally weighted in the scoring process. Some items were worth more points than others. Therefore, once an examiner assigned his or her points on a scale of 10 to 100, those points were then multiplied by the total number of points

possible. That total was multiplied by .01 to convert it to a transformed score. For example, suppose an examiner gave item 1.1 (Organizational Leadership) 40 points. Item 1.1 was worth a maximum of 80 points; 40 points multiplied by 80 points equals 3200 points. The 3200 point total was then multiplied by .01 so that the final score was 32. Once the examiner assigned a score to every item, he or she then added all the converted item scores to generate the total score, which was called the Individual Total Score in this study.

Each examiner was trained and instructed to follow the same process. All score sheets were then sent to team leaders who calculated the average of the individual total scores and recorded it as the grand total points which, for the purpose of this research was called the team overall score. Once this step was complete, examiners coordinated for a consensus meeting. Consensus meetings could occur via telephone conference or face to face. The consensus meetings typically lasted several hours as all examiners had to reach consensus on a score for every item.

During the consensus meeting, individual examiner scores were not part of the conversation since the group was required to come to agreement over each item score. Because each of the examiners had already gone through the process of evaluating the organization and scoring each item individually, they had different viewpoints on the strengths and weaknesses of the organization's processes. Consequently, their individual scores may not have matched each other. The purpose of the consensus meeting was to allow every team member to discuss his or her viewpoint and then come to an agreement on what the group felt the score should be in light of the

discussion. This discussion is typically considered a strength of the scoring process as varied viewpoints are expected to strengthen the accuracy and validity of the score.

Oftentimes during consensus meetings, one examiner will see something in the application that another examiner did not and therefore did not reflect in his or her score. The theory behind the consensus meeting is that by discussing each item extensively as a group and coming to consensus, the team is able to conduct a more thorough analysis and is therefore able to reflect a more consistent and unbiased score for the applicant. It was this ”theory” that was examined in this study. Therefore, the consensus score was used as the score against which examiners were measured when analyzing the data for this study.

Once each item was given a score (using the same 10-point increment and 100-point scale that individual examiners used) the process was repeated to generate a consensus score for each item. For example, remember that item 1.1 was worth a maximum of 80 points. Suppose the examiner team came to a consensus that the organization should be assigned 60 points out of 100. The 60 points were multiplied by 80 points for a total of 4800 points. The 4800 points were then multiplied by .01 to get a score of 48. Each item consensus score was added to create the consensus score for each of the seven categories. Then the category consensus scores were added and the sum was labeled the consensus grand total points. For the purpose of this study, the consensus grand total points is referred to as the Team Consensus Score.

Population

The population for this study included all examiners who completed organizational assessments for the Texas Award for Performance Excellence from 2001 through 2004 (Award Level – Option III), for a total of 34 organizations. All scoring data, including individual scores from each examiner, consensus scores, and overall scores for each organization was included in the data sets provided by the Quality Texas Foundation. Names were removed from applications upon receipt by the Quality Texas Foundation and replaced with codes to ensure anonymity during the scoring process.

Examiners consisted of men and women at various levels of experience in quality management and category and from a variety of professional fields. For the purpose of this study, examiners were grouped into three categories based on their level of experience. The categories included new, returning, and senior level examiners. New examiners were those examiners who were assessing organizations for the first time. Senior examiners were those who had more than one year of experience, or who were in a line of work directly related to quality assessment and had at least one year of experience working with the TAPE organization. Returning examiners were those examiners who were coming back for a second year or who had skipped years in between the current assessment year and their first year of serving as an examiner. There were 250 examiners in this study, including 120 new examiners, 77 returning examiners, and 53 senior examiners.

Each organizational assessment team was made up of a combination of 7 to 9 examiners. Each team included new, returning, and senior examiners. New teams

consisted of greater than 50% new examiners. Senior teams consisted of greater than 50% Returning and Senior examiners. Average teams were split evenly with 50% New examiners and 50% Returning and Senior examiners. Returning and Senior examiners were grouped together based on the fact that both groups had prior experience serving as examiners would have similar influence on the outcome of organizational assessments. The ratio of experience levels varied from team to team based on the number of applicants and number of examiners in a given year. The staff at Quality Texas used the North American Industrial Classification System Code (NAICSC) and their institutional knowledge to create teams that were as qualified yet diverse as possible.

Data Analysis

For the purpose of this study, all data from the Quality Texas Foundation into a database that could interface with the computer program SPSS 14.0. Because the set of TAPE applicants included in the study constitute the entire population of TAPE applicants from 2001 to 2004, descriptive statistics were appropriate for producing informative data that could be analyzed for variation and stability in the scoring process. Exploration of patterns in descriptive statistics and multivariate analysis of variance (MANOVA) were the primary tools used in this particular study, along with Cronbach's Alpha as an indicator of reliability.

Dependent Variable Generation

Since scoring for the TAPE is based on an individual examiner's best subjective assessment of how well an organization meets the criterion for various items, it was impossible to have one objective score against which all other scores could be measured. Therefore, the team consensus score was used as the score against which individual examiner scores were measured.

One of the first steps in analyzing the data to answer the research questions was to establish deviation scores. Deviation scores were calculated by using the consensus scores as the anchor point. Therefore, analyses conducted for Research Question 1, which focused on overall results, were obtained by subtracting the total consensus score from each of the individual total scores to obtain individual total deviation scores. Individual total deviation scores were then averaged to get a team mean deviation score for each of the 34 teams.

Analysis for Research Question 2, which was designed to reveal individual examiner information on separate items was conducted by subtracting team item consensus scores from individual item scores in each category resulting in deviation scores for each examiner for each of the 17 items studied. Item deviation scores were then averaged by item to get an item mean deviation score for each item for each of the 34 teams.

Categories and Items

Although the Texas Award for Performance Excellence has always consisted of seven main categories including Leadership, Strategic Planning, Customer and

Market Focus, Measurement, Analysis, and Knowledge Management, Human Resource Focus, Process Management and Business Results, there have been variations from year to year on the number of items within each category. As a result of such variation and the fact that data collected for this study extends across four years, some items contained incomplete data and therefore were not used. Results for Cronbach's Alpha indicated that removing three incomplete items from the data set did not affect the overall reliability of the data. Therefore, 17 items that were common across the time period were tested in the analysis for each of the four research questions. A description of each item is displayed in Table 8. This information was taken from the 2005 Criteria for Performance Excellence for the Texas Award for Performance Excellence.

TABLE 8. Description of Items From the TAPE Criteria for Performance Excellence

Category 1.0 Leadership	
Item 1.1 Senior Leadership	This item focuses on how senior leaders guide and sustain the organization, how they communicate with employees and how they encourage high performance.
Item 1.2 Governance and Social Responsibility	This item focuses on how the organization addresses its responsibilities to the public, how it ensures ethical behavior and what it does to practice good citizenship.

TABLE 8. Continued

Category 2.0: Strategic Planning	
Item 2.1 Strategy Development	This item addresses how the organization develops strategic objectives and action plans and how they are measured.
Item 2.2 Strategy Deployment	This item focuses on how the organization converts the strategic objectives and action plans and also on what performance measures or indicators are developed based on the objectives and action plans.
Category 3.0: Customer and Market Focus	
Item 3.1 Customer and Market Knowledge	This item asks how the organization determines requirements, expectations and preferences of customers and markets to ensure the continuing relevance of the organization's products and services and to develop new opportunities.
Item 3.2 Customer Relationships and Satisfaction	This item addresses strategies that the organization uses to build relationships in order to acquire, satisfy and retain customers, increase customer loyalty and develop new opportunities.
Category 4.0: Measurement, Analysis, and Knowledge Management	
Item 4.1 Measurement, Analysis, and Review of Organizational Performance	This item focuses on how the organization measures, analyzes, aligns, reviews, and improves it performance throughout the organization.
Item 4.2 Information and Knowledge Management	This item focuses on how an organization ensures quality and availability of needed data for employees, suppliers, partners and customers.

TABLE 8. Continued

Category 5.0: Human Resource Focus	
Item 5.1 Work Systems	This item addresses how the organization's work and jobs enable employees and the organization to achieve high performance through compensation, career progression and related workforce practices.
Item 5.2 Employee Learning and Motivation	This item addresses how the organization's employee education, training and career development support the achievement of overall objectives and contribute to high performance for the organization.
Item 5.3 Employee Well-Being and Satisfaction	This item addresses how the organization maintains a work environment and an employee support climate that contributes to the well-being, satisfaction and motivation of all employees.
Category 6.0: Process Management	
Item 6.1 Value Creation Processes	This item focuses on how the organization identifies and manages its key processes for creating customer value and achieves its business success and growth.
Item 6.2 Support Processes and Operational Planning	This item focuses on how the organization manages its key processes that support value creation, financial management and continuity of operations in an emergency.
Category 7.0: Business Results	
Item 7.1 Product and Service Outcomes	This item considers the organization's summary of overall key product and service performance results.
Item 7.2 Student- and Stakeholder-Focused Results	This item considers the organization's summary of overall key customer-focused results including customer satisfaction and perceived value.
Item 7.3 Financial and Market Results	This item considers the organization's summary of key financial and marketplace performance results.
Item 7.4 Human Resource Results	This item considers the organization's summary of overall key human resource results, including work system performance and employee learning, development, well-being and satisfaction.

Reliability

Reliability of the scoring process was established using Cronbach's Alpha. Twenty items were embedded within the seven categories of the criterion. However, data for three of the twenty items was incomplete which made it impossible to run a test for all items. Therefore three separate coefficients were calculated; the first test was run with one of the incomplete items removed; a second test was run with two of the three items removed; a third test was run with all three incomplete items removed. The alpha score in all three tests was .940 or higher indicating that, even with the removal of incomplete items, the data was stable and reliable. A description and summary table of the coefficient calculations is included in Chapter IV (Table 9).

Research Question 1

Research Question 1 was, "Is the mean of the deviations of individual total scores from total consensus score equal to zero?" This question was addressed through the analysis of descriptive statistics, histograms and other graphs. Since total consensus scores were considered the true score for each of the organizational assessments, the closer the mean of the deviations of individual total scores was to zero, the more consistent in scoring the team was considered to be. For the purpose of this question, the mean of the individual total deviation scores was called the team mean deviation score. In order to determine the team mean deviation score, the total consensus score was subtracted from each of the individual total scores to get an individual total deviation score for each examiner on a team. Next, the mean of the individual total

deviation scores were calculated for each of the 34 teams. An analysis was run for all four years of data combined and by each separate year.

Research Question 2

As in question one, Research Question 2 was, “Is the mean of the deviations of individual item scores from team item consensus scores equal to zero?” This question was addressed through the analysis of descriptive statistics, histograms and other tables. However, for the purpose of question two, individual item scores and team item consensus scores were used. Since team item consensus scores were considered to be the true scores for each item, the closer the mean deviation for individual item scores was to zero, the more consistent in scoring the examiners were considered to be. An analysis was run for each of the 17 items embedded in the Criteria for Performance Excellence by using descriptive statistics.

Research Question 3

Research Question 3 was, “Do item deviation scores vary across the following classifications:

a. Levels of Examiner Experience
b. Sectors
c. Levels of Self-Assessment
d. Levels of Team Experience

In question three, a cross tabulation and multivariate analysis of variation (MANOVA) were used. A cross-tabulation table of individual rater experience and

sector were created to establish whether or not the cell sizes were sufficiently large. Results of the cross-tabulation revealed that cells sizes were too small to compare variables across the levels. Consequently, raw data were loaded into the statistical program SPSS and deviation scores were calculated to allow for observation within each classification factor. A MANOVA was conducted to determine if there was a significant difference across the independent variables (levels of experience, sector, levels of self-assessment and levels of team experience). Next, Univariate F tests (ANOVA) were conducted to determine if there was variation within the independent variables and then Post Hoc tests were conducted to determine the location of differences across the independent variables.

Additional Description of Data

The following line charts (Figures 1-17) represent additional description of the data in terms of how the item mean deviation scores fall across sectors and years by each item. Because this study is the first attempt to analyze authentic data from TAPE, it is necessary to make sure that the data has been observed and described from many angles. While there may appear to be patterns in some of charts or across some items, it is not possible to know what caused the pattern or if, in fact, it does represent a pattern. Possibilities for anomalies may result from the applicant organizations' experience level in quality strategies or from the applicant organizations' abilities to fill out the organizational profile and application in such a way that enables examiners to be able to see the same strengths and opportunities for improvement. Anomalies or patterns may result from particularly experienced or

inexperienced teams or from the nature of difficulty involved in assessing a particular item. Since neither the organizations nor teams of examiners repeat, it is not possible to compare from year to year or from sector to sector. However, observing the overall picture is helpful in gaining an understanding of the data and possibly generating questions for further research.

Average of Item Mean Deviation Scores for Item 1.1

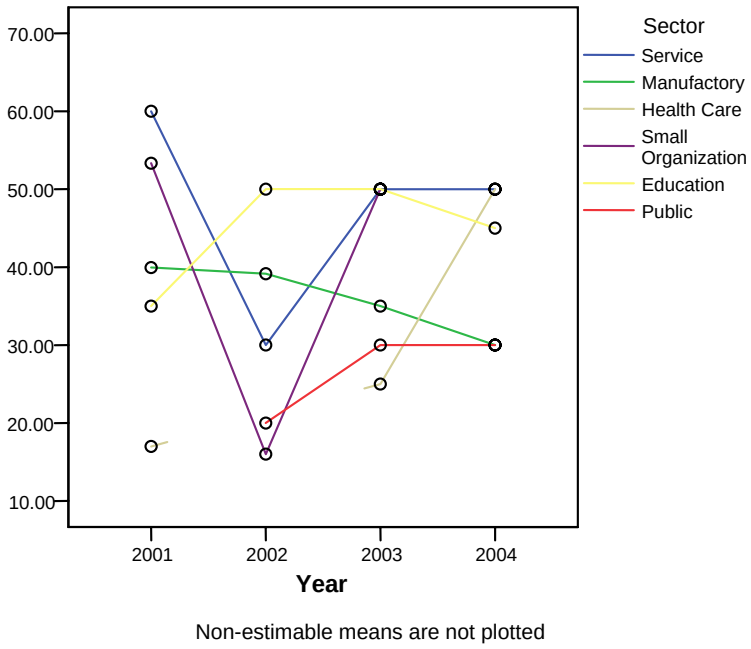

FIGURE 1. Average of Item Mean Deviation Scores for Item 1.1

Average of Item Mean Deviation Scores for Item 1.2

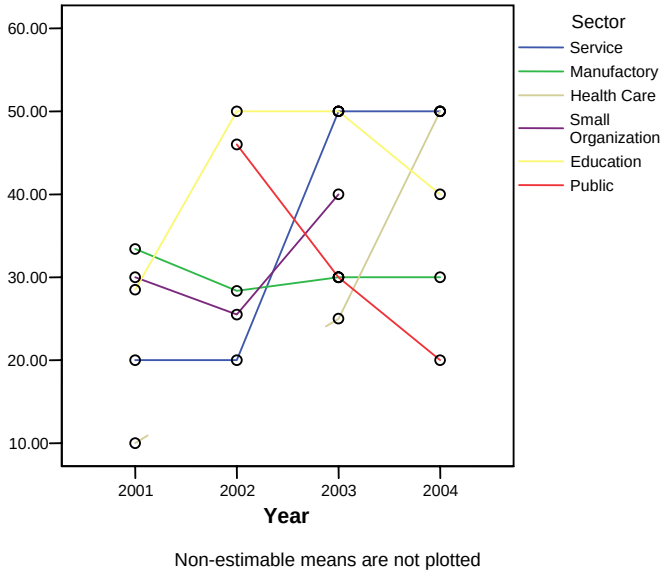

FIGURE 2. Average of Item Mean Deviation Scores for Item 1.2

Average of Item Mean Deviation Scores for Item 2.1

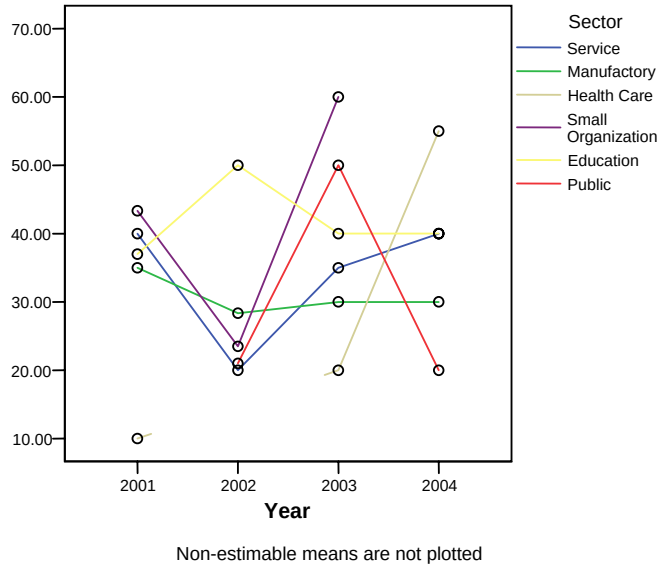

FIGURE 3. Average of Item Mean Deviation Scores for Item 2.1

Average of Item Mean Deviation Scores for Item 2.2

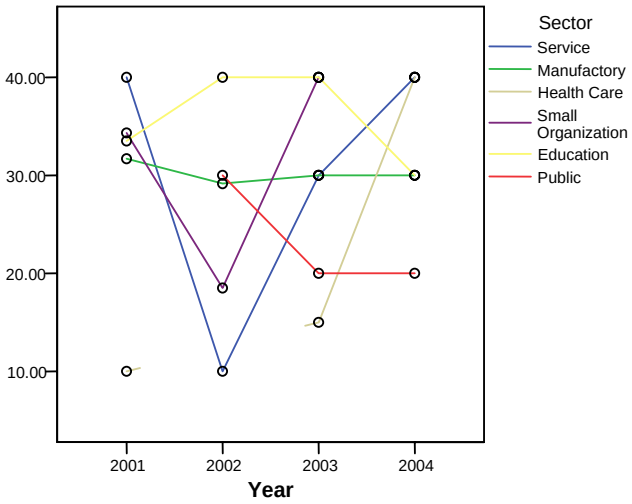

FIGURE 4. Average of Item Mean Deviation Scores for Item 2.2

Average of Item Mean Deviation Scores for Item 3.1

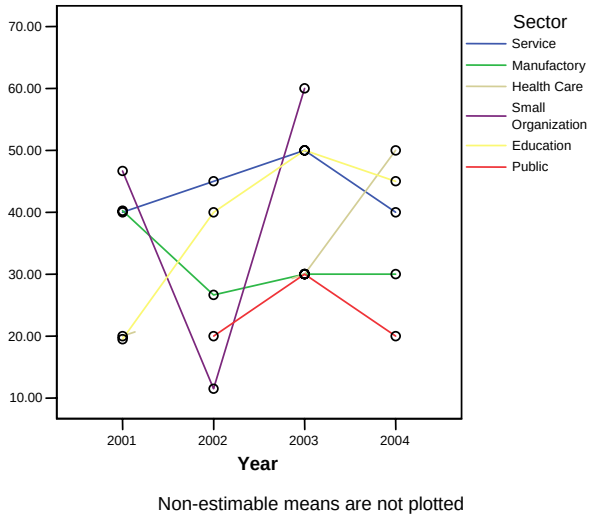

FIGURE 5. Average of Item Mean Deviation Scores for Item 3.1

Average of Item Mean Deviation Scores for Item 3.2

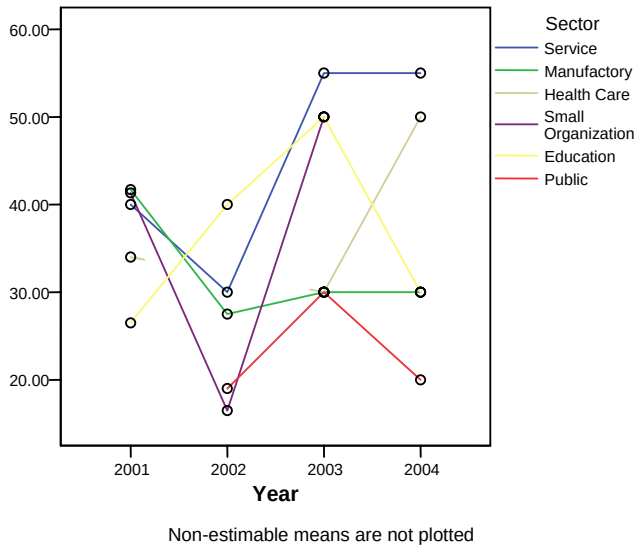

FIGURE 6. Average of Item Mean Deviation Scores for Item 3.2

Average of Item Mean Deviation Scores for Item 4.1

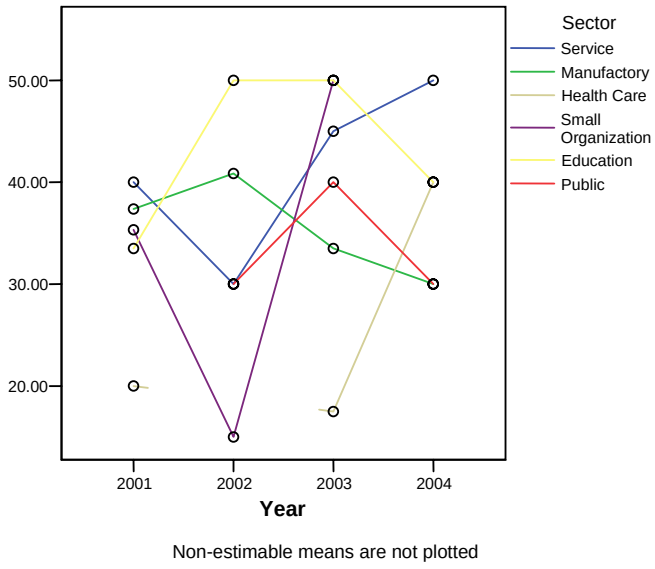

FIGURE 7. Average of Item Mean Deviation Scores for Item 4.1

Average of Item Mean Deviation Scores for Item 4.2

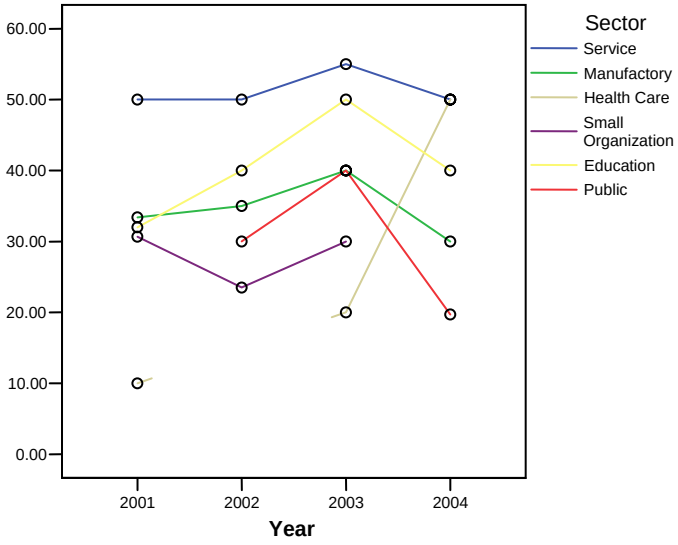

FIGURE 8. Average of Item Mean Deviation Scores for Item 4.2

Average of Item Mean Deviation Scores for Item 5.1

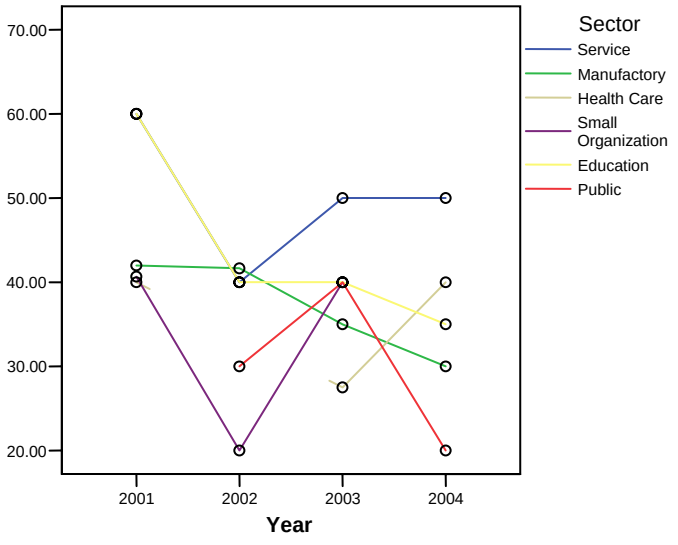

FIGURE 9. Average of Item Mean Deviation Scores for Item 5.1

Average of Item Mean Deviation Scores for Item 5.2

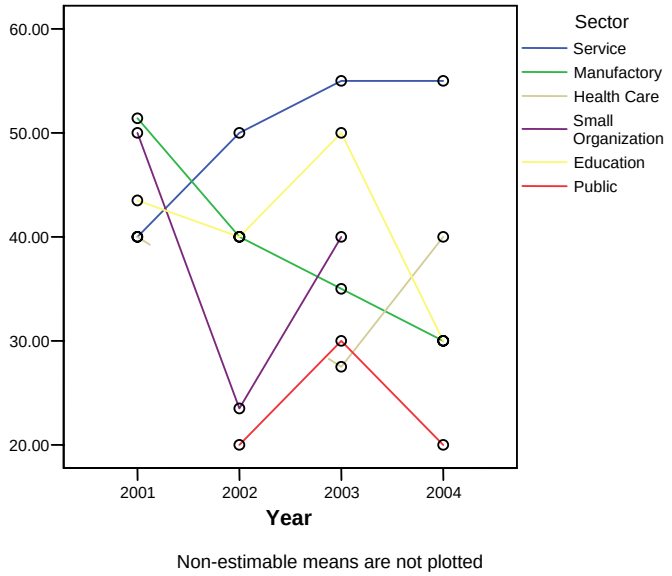

FIGURE 10. Average of Item Mean Deviation Scores for Item 5.2

Average of Item Mean Deviation Scores for Item 5.3

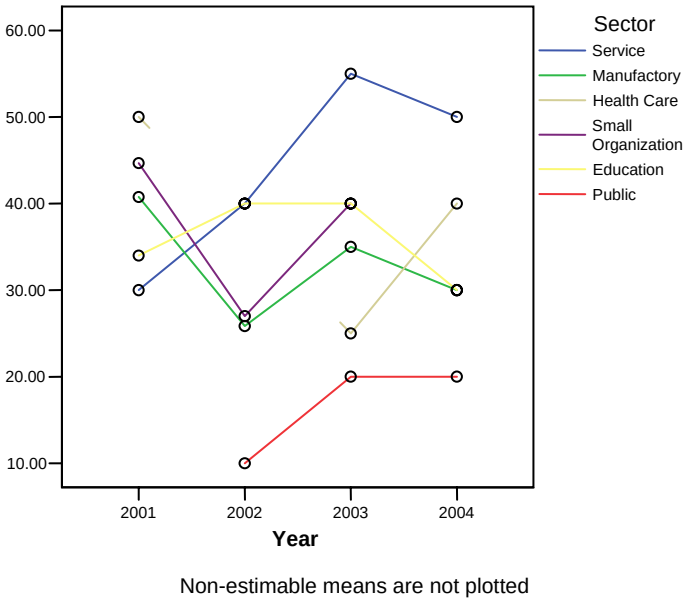

FIGURE 11. Average of Item Mean Deviation Scores for Item 5.3

Average of Item Mean Deviation Scores for Item 6.1

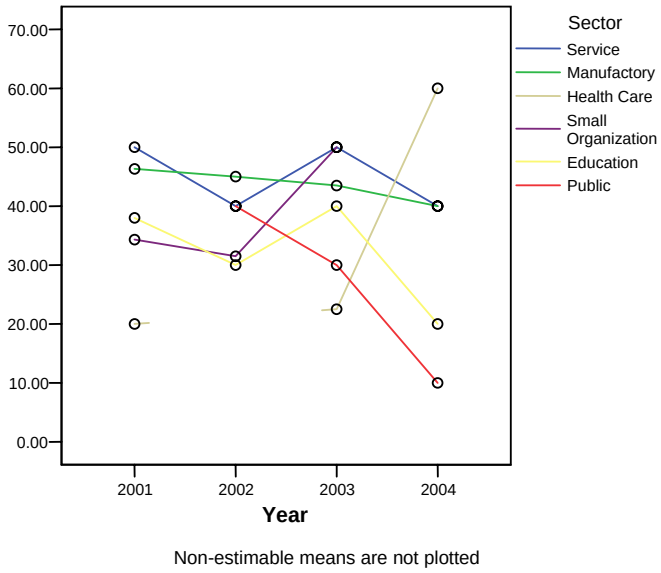

FIGURE 12. Average of Item Mean Deviation Scores for Item 6.1

Average of Item Mean Deviation Scores for Item 6.2

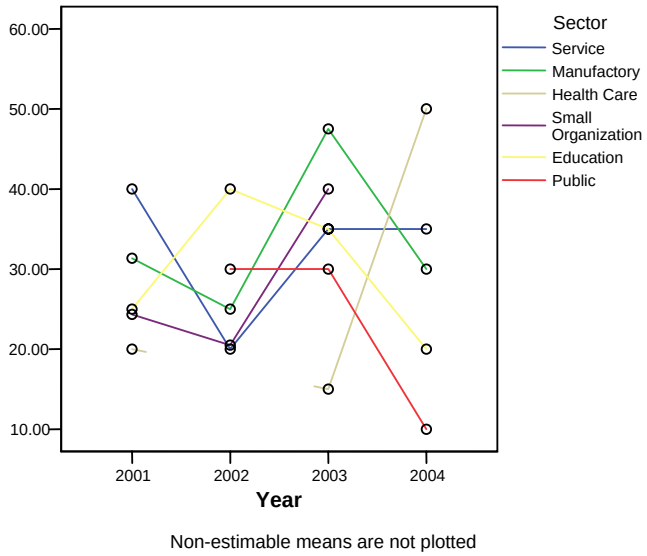

FIGURE 13. Average of Item Mean Deviation Scores for Item 6.2

Average of Item Mean Deviation Scores for Item 7.1

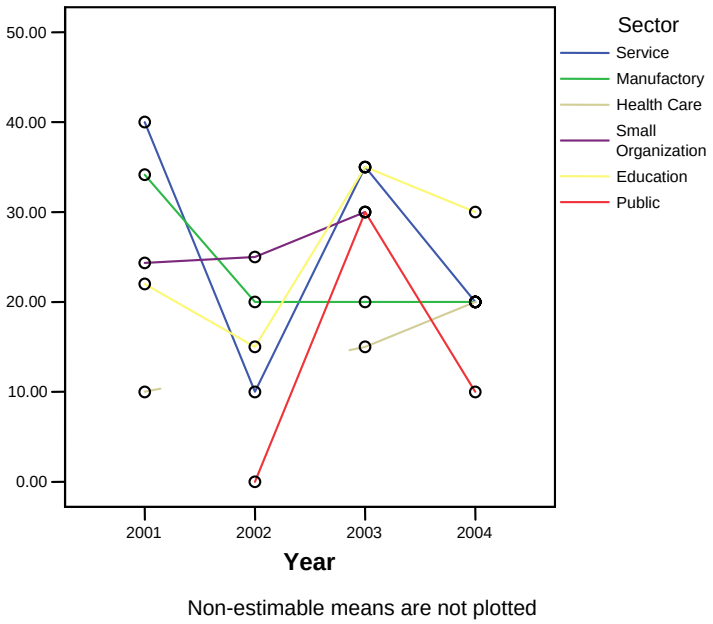

FIGURE 14. Average of Item Mean Deviation Scores for Item 7.1

Average of Item Mean Deviation Scores for Item 7.2

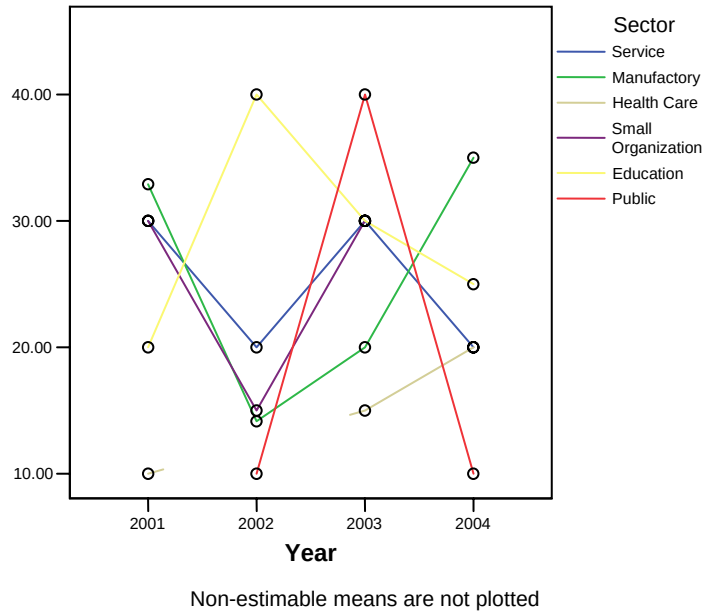

FIGURE 15. Average of Item Mean Deviation Scores for Item 7.2

Average of Item Mean Deviation Scores for Item 7.3

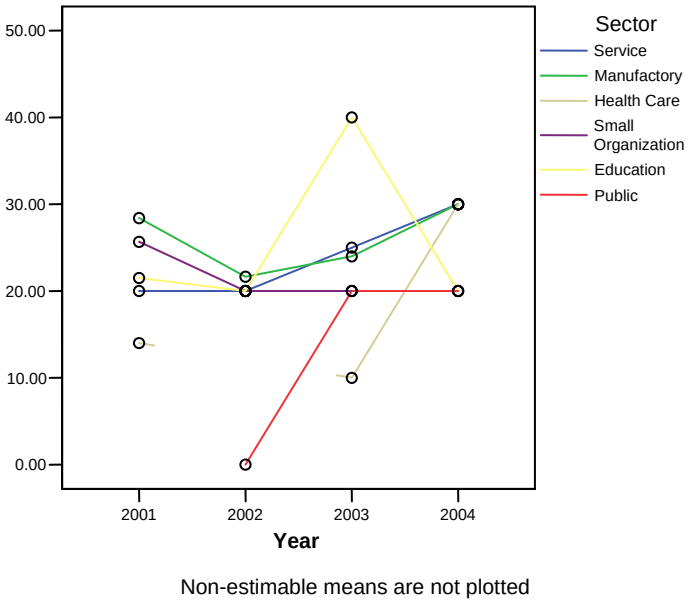

FIGURE 16. Average of Item Mean Deviation Scores for Item 7.3

Average of Item Mean Deviation Scores for Item 7.4

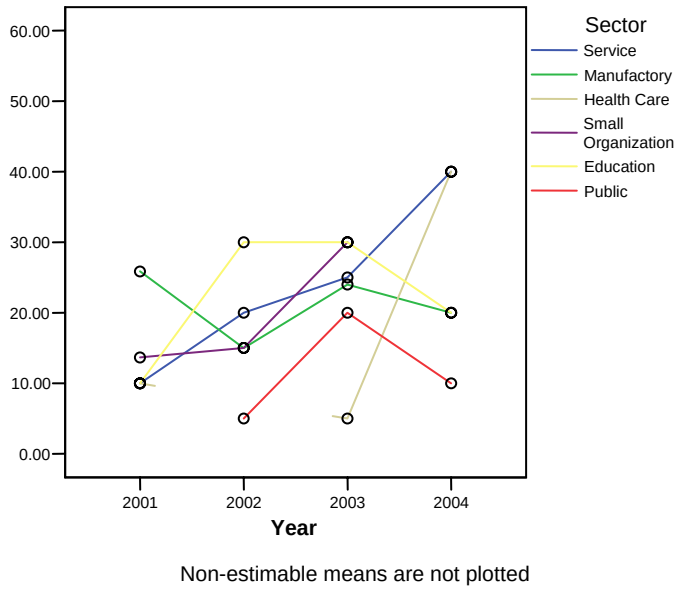

FIGURE 17. Average of Item Mean Deviation Scores for Item 7.4

Summary

The purpose of this study was to determine the scoring stability of third-party examiners who were assessing organizational performance for the Texas Award for Performance Excellence. All applications and associated examiner scoring data for the TAPE from 2001 through 2004 were collected allowing for appropriate descriptive statistics to be used in analysis for each research question. Analysis of these data is presented in Chapter IV.

CHAPTER IV

ANALYSIS OF DATA

The results of the data analyses on a question by question basis are detailed in Chapter IV. Raw data obtained from the Quality Texas Foundation and converted to a usable database are contained in Appendix A, while summary statistics and graphical representations of the data are presented in the tables and figures following each research question. The purpose of this study was to determine the scoring stability of third-party examiners who scored organizational assessments for the Texas Award for Performance Excellence. The first two research questions were addressed using descriptive statistics. The last research question represents further analyses through the use of MANOVA and post hoc tests.

Reliability

Tests for reliability of the scoring process were run using Cronbach's Alpha. Overall, there were 20 items spread across the seven categories. As noted in Chapter III, tests for reliability using all 20 variables could not be executed because there were too few cases for the analysis for three of the variables. Table 9 is a summary of tests using 19, 18, and 17 items where Items 6.3, 7.5, and 7.6 were removed due to incomplete data.

The observer will see in Table 9 that, in all cases, the obtained reliability coefficients were excellent regardless of instances where there were incomplete responses. Even when more than 80% of the data was removed, the reliability of these

scores showed relatively little change, indicating that these scores were very consistent and stable.

TABLE 9. Summary of Reliability Coefficient Calculations

Item 6.3 Removed		
Cases	**N**	**%**
Valid	47	18.8
Excluded (a)	203	81.2
Total	250	100.0
Cronbach's Alpha = .942		
Item 6.3 and 7.5 Removed		
Cases	**N**	**%**
Valid	47	18.8
Excluded (a)	203	81.2
Total	250	100.0
Cronbach's Alpha = .940		
Item 6.3, 7.5 and 7.6 Removed		
Cases	**N**	**%**
Valid	250	100.0
Excluded (a)	0	0
Total	250.	100.0
Cronbach's Alpha = .959		

Research Questions

This study was an analysis of data collected by the Quality Texas Foundation to determine the scoring stability of third-party examiners when assessing organizational performance for the Texas Award for Performance Excellence. The research was conducted through exploring the following three questions:

1. Is the mean of the deviations of individual total scores from team total consensus score equal to zero?

2. Is the mean of the deviation of individual item scores from team item consensus scores equal to zero?
3. Do item deviation scores vary across the following classifications:
    a. Levels of examiner experience
    b. Sector
    c. Levels of self assessment
    d. Levels of team experience

Research Question 1 – Is the mean of the deviations of individual total scores from team total consensus score equal to zero?

Research Question 1 was addressed through the development and analysis of descriptive statistics, histograms and other graphs. Since total consensus scores were considered the true score for each of the organizational assessments, the closer the mean of the deviations of individual total scores was to zero, the more consistent in scoring the team was considered to be. For the purpose of this question, the mean of the individual total deviation scores was called the team mean deviation score. In order to determine the team mean deviation score, the total consensus score was subtracted from each of the individual total scores to get an individual total deviation score for each examiner on a team. Next, the mean of the individual total deviation scores was calculated for each of the 34 teams. An analysis was run for all four years of data combined and by each separate year.

Descriptive statistics results for 2001-2004. The team mean deviation for total consensus scores for all four years of data combined was 96.96 with a standard deviation of 58.7. The skewness was -.249. In general, individual examiners did not produce scores that were consistent with the team total consensus scores. A summary of the descriptive statistics results are presented in TABLE 10.

TABLE 10. Summary Statistics for Team Mean Deviation Scores for 2001-2004

Description	Statistics
Mean	96.9597
Std. Deviation	58.73019
Minimum of	204.58
Maximum of	-54.07
Skewness	-.249
Kurtosis	-.351

There were 250 examiners, each with an individual total score. An individual total deviation score was generated for each examiner by subtracting the total consensus score from the individual total score. The individual total deviation scores were then averaged to produce a team mean deviation score. Examiners who were on the same team each shared the same team mean deviation score. Figure 18 is a summary of the frequency with which the 250 individual examiners produced the same team mean deviation score between the years 2001 and 2004.

A summary of the frequency of *teams* who shared the same team mean deviation scores is presented in Figure 1. Because each team had several examiners, the examiners shared the same team mean deviation score and therefore the team only

produced one observation with regard to team mean deviation score. In Figure 1 each *examiner's* team mean deviation score was used as a separate observation and, therefore is a representation of how individual examiners from one team cluster with individual examiners on different teams that had the same team mean deviation score.

For this study, the closer the team mean deviation score is to zero, the more consistent examiners are considered to be. Therefore, scores above zero can be interpreted as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be interpreted as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

Looking at all examiners over the entire four years, the skewness value of -.249 indicates that scores are shifted above zero. The mean for the distribution is 96.96 which, indicates that examiners gave overall scores that were higher or more lenient than scores produced as a result of coming to team consensus. There also appears to be considerable variation among the team mean deviation scores, which is evident by the range of a minimum of -54 and a maximum of 204. Results from Research Questions 2 and 3 will serve to analyze further this variation.

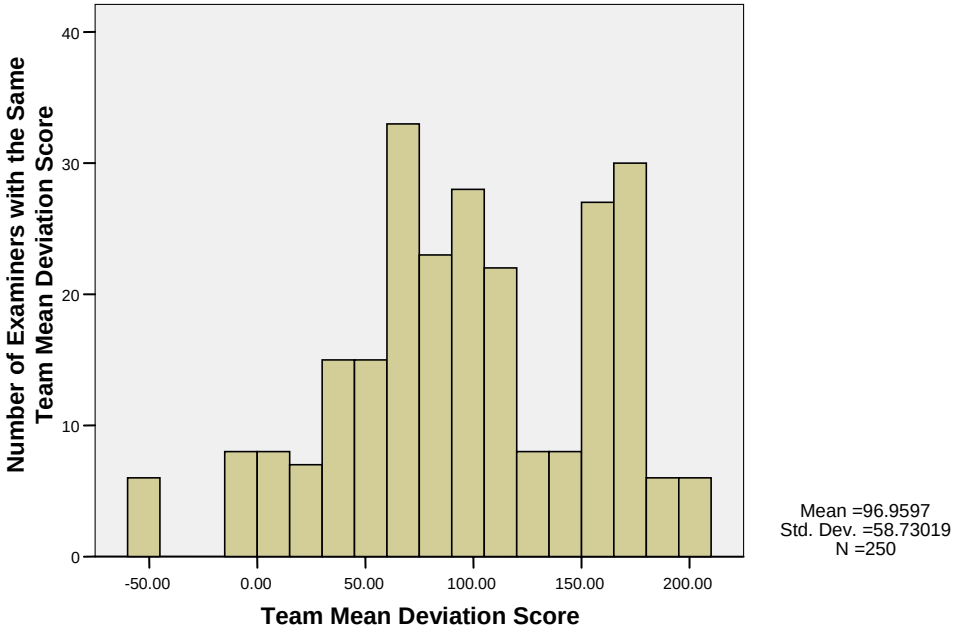

FIGURE 18. Distribution of the Frequency of Examiners' Team Mean Deviation Scores in Regard to Total Score

The total team deviation scores for each of the 34 teams are presented in Figure 18. This figure represents the same data as in Figure 1 except that the 250 examiners are grouped by their teams rather than by other examiners on teams with the same team mean deviation score. Whereas Figure 18 was a summary of the number of examiners who had the same team mean deviation scores in regard to total scores, Figure 19 is a summary of how far each unique team's mean deviation score was from that team's total consensus score. Again, the closer the team mean deviation score was to zero, the more consistent that team's examiners' scores were considered to be. Therefore, by observing Figure 19, it can be ascertained that, with the exception

of teams 15, 25 and 32, examiners did not score consistently with the total consensus scores and, in general, tended to be more lenient overall.

Team 11 appears to be a major outlier. Investigation of the raw data revealed that only six of the seven examiners had recorded scores. Consequently, the team mean deviation score was calculated using only six sets of data, but was averaged based on seven scores. This skewed the results for that unique case and should not be considered when analyzing the data in Figure 19.

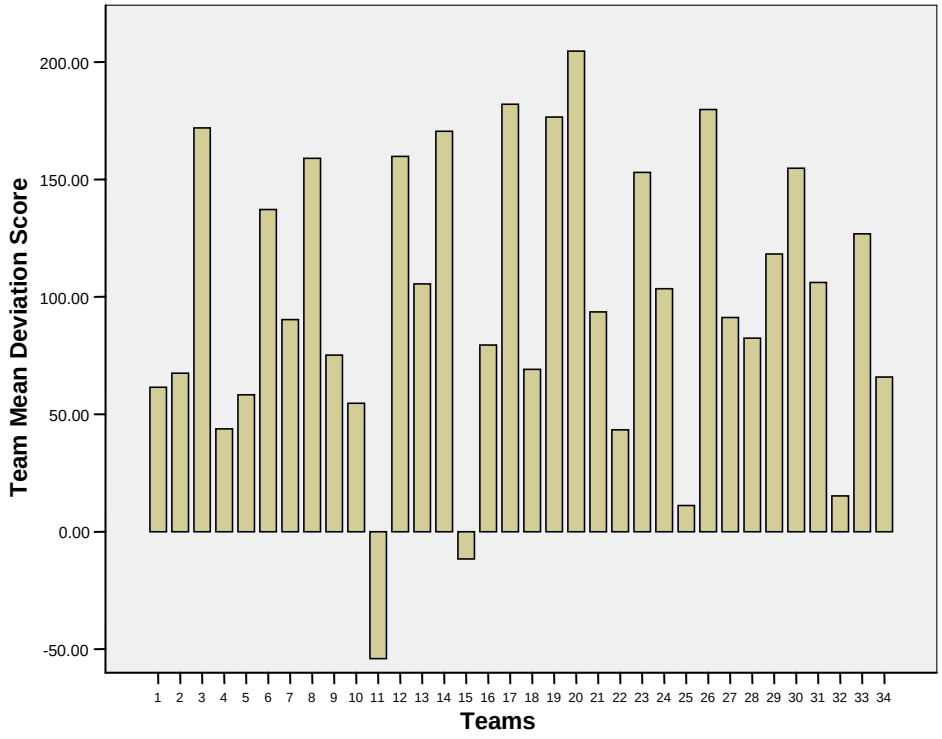

FIGURE 19. Distribution of Team Mean Deviation Scores by Teams

Teams 15, 25 and 32 are the three teams whose team mean deviation scores were closest to zero or were most like the total consensus scores for their respective teams. All three teams were senior-level teams in terms of overall experience. Senior-level teams are those teams that are made up of more than 50% senior and returning examiners. Team 15 assessed an educational organization in 2002. Team 25 assessed a service organization in 2003. Team 32 assessed a health care organization in 2004. Based on this initial observation, the only pattern appears to be that of level of team experience. Each of the three teams assessed organizations in different sectors and in different years, so it would appear that the level of team experience was a consistent and possibly influencing factor. However, in observing those teams whose team mean deviation scores were farthest away from their total consensus score, results appear to be similar in that all teams had a senior level of experience.

Teams 3, 17, 20 and 26 appear to have the greatest distance from zero or had examiners who collectively, were least like the total consensus scores. Three of the four teams were senior-level teams in terms of overall experience. Team 20 was an average team in terms of overall team experience, meaning that the team consisted of 50% New examiners and 50% Returning and Senior examiners. Team 3 assessed a health care organization in 2001. Team 17 assessed an educational organization in 2002. Team 20 assessed a public organization in 2003. Team 26 assessed an educational organization in 2003. The only repeated pattern here is that three of the four teams were of senior-level experience, two of the four teams assessed educational organizations and two of the four teams assessed organizations in 2003. In general, there does not appear to be a consistent pattern. However, additional

analysis will be conducted in Research Questions 2 and 3 to analyze further these results.

Descriptive statistics results for 2001. There were 87 examiners in 12 teams for 2001. Considering that the closer the team mean deviation score is to zero, the more consistent examiners are considered to be, then scores above zero can be translated as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be translated as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

The team mean deviation for 2001 was 85.45 with a standard deviation of 58.7. The skewness was -.385. In general, examiners did not produce individual total scores that were consistent with their team's total consensus score. A summary of the descriptive statistics results are presented in Table 11.

TABLE 11. Summary Statistics for Team Mean Deviation Scores for the Texas Award for Performance Excellence in 2001

Description	Statistics
Mean	85.4474
Standard Deviation	58.69183
Minimum of	-54.07
Maximum of	171.93
Skewness	-.385
Kurtosis	.200

The 87 examiners were grouped into 12 teams. Each examiner had an individual total score which was used to generate an individual total deviation score from their team's total consensus score. These individual total deviation scores were then averaged to obtain a team mean deviation score for each team. Figure 20 is a summary of the number of examiners who were on teams that had a particular team mean deviation score. The team mean deviation for the distribution in 2001 was 85.45, which indicates that examiners in 2001 scored more leniently when working alone than when coming to consensus as a team. It is important to note that the one outlier in this data set is the previously mentioned team (Team 11) that had a team mean deviation score based on incomplete data. The standard deviation for teams in 2001 was 58.69 and the range fell across a minimum of -54.07 and a maximum of 171.93 which indicated large variation. It is important to remember that this variation was impacted by the outlier team, although when eliminating this team's variation, the remaining teams' variation is still considerable.

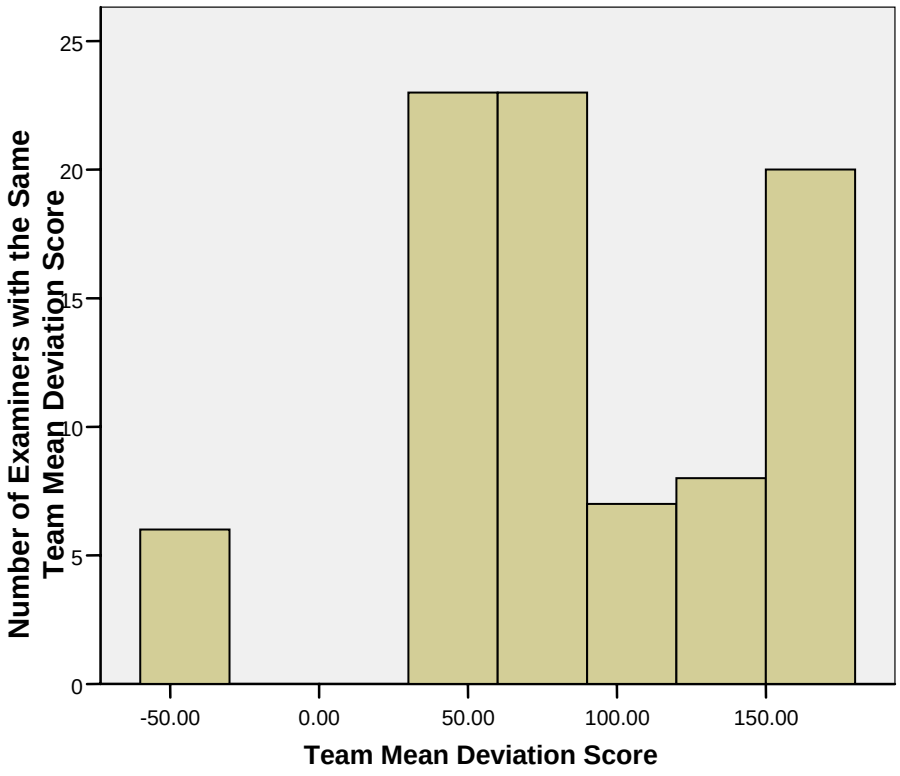

FIGURE 20. Distribution of the Frequency of Examiners' Team Mean Deviation Scores in Regard to Total Score in 2001

The total team deviation scores for each of the 12 teams is presented in Figure 21. Whereas Figure 20 was a summary of the number of examiners who were on teams that had a particular team mean deviation score, Figure 21 is a summary of how far each unique team's mean deviation score was from that team's total consensus score.

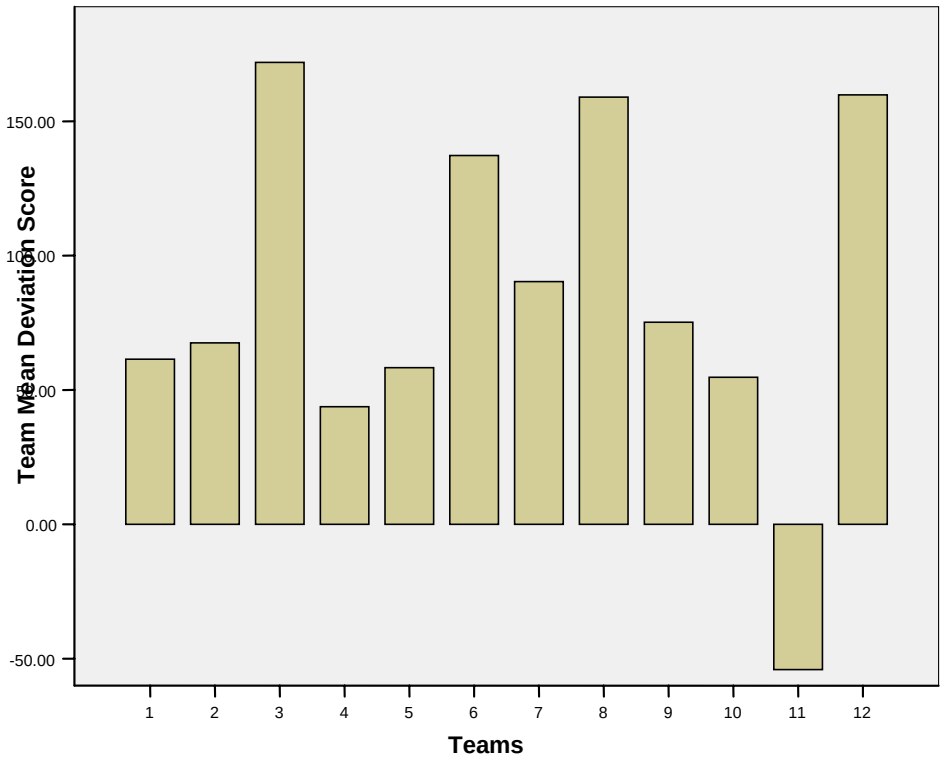

FIGURE 21. Distribution of Team Mean Deviation Scores by Team in 2001

Again, the closer the team mean deviation score was to zero, the more consistent that team's examiners' scores were considered to be. By observing the bars in the histogram, one can easily ascertain that examiners did not produce team mean deviation scores that were consistent with the total consensus score. Teams 3, 6, 8, and 12 appear to have the greatest variation compared to the total consensus score for their team. Team 3 was a senior-level team (meaning that 51% of the team was made up of senior and returning examiners) that assessed a health care organization. Team 6 was an average-level team (meaning that 50% of the team were senior and returning

examiners and 50% of the team were new examiners) that assessed a manufacturing organization. Team 8 was a new-level team (meaning that 51% of the team were new examiners) that assessed an educational organization. Team 12 was a senior-level team that assessed a small organization. Overall, there appears to be no particular pattern between or among teams who appear to be inconsistent with their team's total consensus score.

Descriptive statistics for 2002. There were 53 examiners in 7 teams for 2002. Considering that the closer the team mean deviation score is to zero, the more consistent examiners are considered to be, then scores above zero can be translated as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be translated as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

The team mean deviation was 106.99 with a standard deviation of 67.2. The skewness was -.446. In general, examiners did not produce individual total scores that were consistent with their team's total consensus scores. A summary of the descriptive statistics is presented in Table 12.

TABLE 12. Summary Statistics for Team Mean Deviation Scores for the Texas Award for Performance Excellence in 2002

Description	Statistics
Mean	106.9906
Variance	4515.770
Std. Deviation	67.19948
Minimum of	-11.63
Maximum of	182.00
Skewness	-.446
Kurtosis	-.952

Each of the 53 examiners were grouped into 7 teams. Each examiner had an individual total score which was used to generate an individual total deviation score from their team's total consensus score. These individual total deviation scores were then averaged to obtain a team mean deviation score for teach team. Figure 22 is a summary of the number of examiners who were on teams that had a particular team mean deviation score. The team mean deviation in 2002 was 106.99, which indicates that examiners in 2002 scored more leniently when working alone than when coming to consensus as a team and, in general they scored more leniently than examiners in 2001. The standard deviation for teams in 2002 was 67. 20 and the range fell across a minimum of -11.63 a maximum of 182.00 which indicated a large variation.

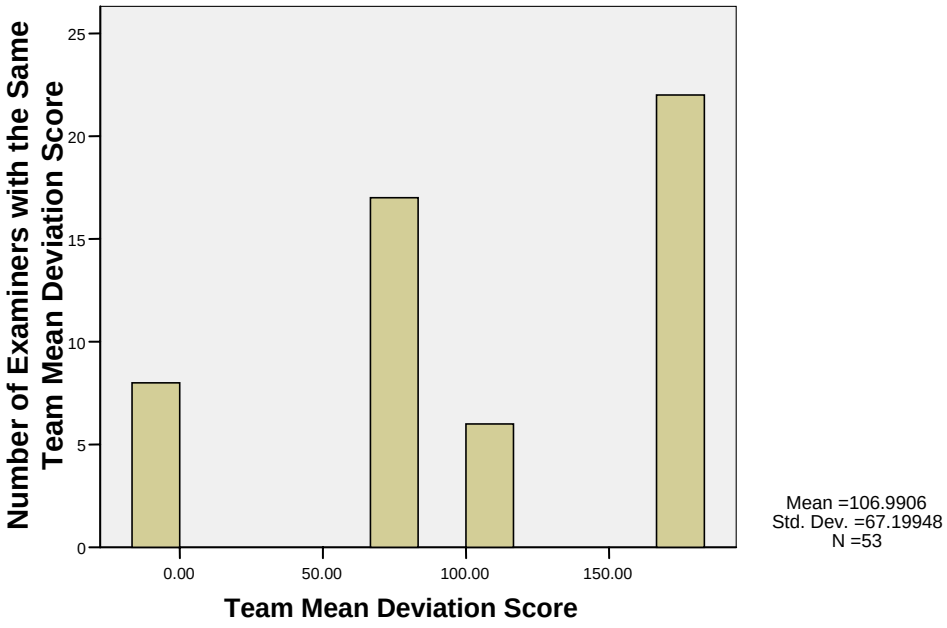

FIGURE 22. Distribution of the Frequency of Examiners' Team Mean Deviation Scores in Regard to Total Score in 2002

The total team deviation scores for each of the 7 teams is presented in Figure 23. Whereas Figure 22 was a summary of the number of examiners who were on teams that had a particular team mean deviation score, Figure 23 is a summary of how far each unique team's mean deviation score was from that team's total consensus score. Again, the closer the team mean deviation score was to zero, the more consistent that team's examiners' scores were considered to be. By observing the bars in the histogram, one can easily ascertain that, in general, examiners did not produce team mean deviation scores that were consistent with the total consensus score.

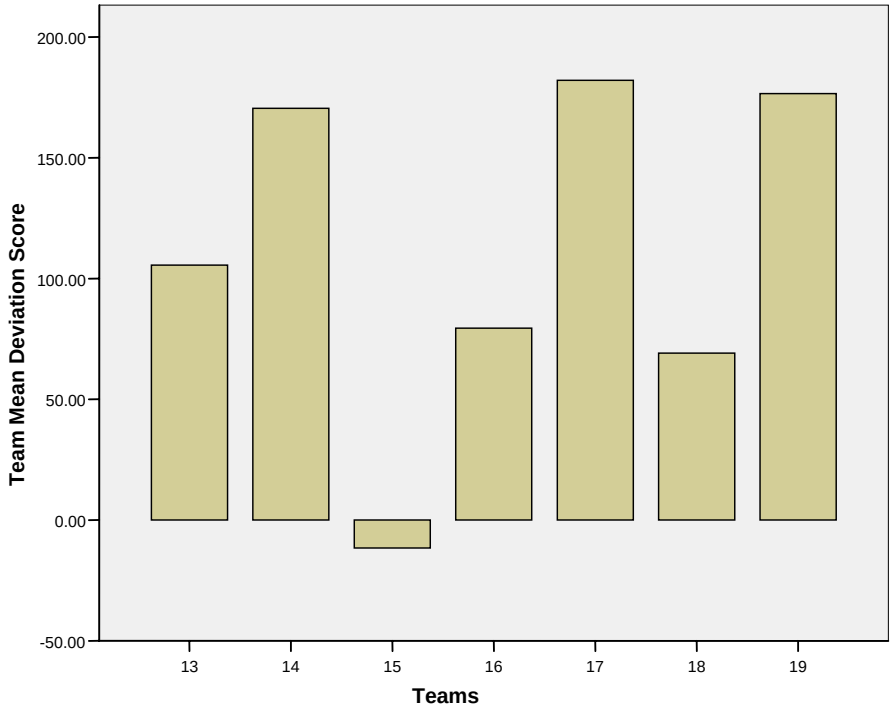

FIGURE 23. Distribution of Team Mean Deviation Scores by Teams in 2002

It does appear that there was some consistency in variation between teams 14, 17 and 19. All three teams had team mean deviations that appear to be substantially inconsistent with their teams' total consensus score. Team 14 was a new team (more than 50% new examiners) that assessed a manufacturing organization. Team 17 was a senior-level team (more than 50% senior and returning examiners) that assessed a small organization. Team 19 was a senior-level team that assessed a service organization. Team 15 was a senior-level team (greater than 50% senior and/or returning examiners) in terms of experience and assessed an educational organization. Team 15 also had significantly less variation than the other teams in 2002. This may have been

due to any number of variables including training, team leadership, experience of the examiners, knowledge of the type of organization, quality of the application, etc.), but due to small cell size, it is not possible to isolate the cause and effect.

Descriptive statistics for 2003. There were 63 examiners in 9 teams for 2003. Considering that the closer the team mean deviation score is to zero, the more consistent examiners are considered to be, then scores above zero can be translated as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be translated as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

The team mean deviation for 2003 was 103.87 with a standard deviation of 59.25. The skewness was .160. In general, examiners did not produce individual total scores that were consistent with their team's total consensus scores. A summary of the descriptive statistics are presented in Table 13.

TABLE 13. Summary Statistics for Team Mean Deviation Scores for the Texas Award for Performance Excellence in 2003

Description	Statistics
Mean	103.8651
Std. Deviation	59.24787
Minimum of	11.13
Maximum of	204.58
Skewness	.160
Kurtosis	-.875

The 63 examiners were grouped into 9 teams. Each examiner had an individual total score which was used to generate an individual total deviation score from their team's total consensus score. These individual total deviation scores were then averaged to obtain a team mean deviation score for each team. Figure 24 is a summary of the number of examiners who were on teams that had a particular team mean deviation score. The team mean deviation for the distribution in 2003 was 103.87, which indicates that examiners in 2003 scored more leniently when working alone than when coming to consensus as a team and, in general they had similar mean deviation scores, but a somewhat different distribution than examiners in 2002. The standard deviation for teams in 2003 was 59.25 while the standard deviation for examiners in 2002 was 67.2. The range fell across a minimum of 11.13 and a maximum of 204.58 which indicated a large variation. The skewness value for team mean deviations scores was .160 while the skewness values for 2002 was -.446. 2003 was the only year out of the four years with a skewness value that was in the positive range. All of the team mean deviations fell in a range above zero which makes 2003 data different from the previous two years and indicates that all examiners in all teams scored more leniently than the total consensus score.

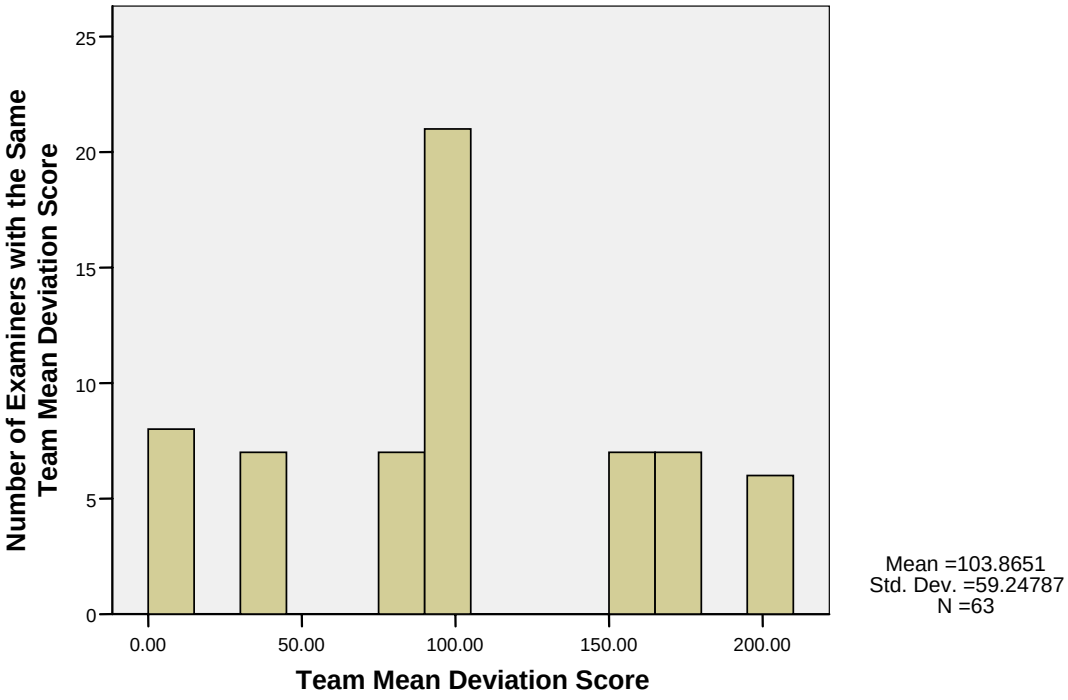

FIGURE 24. Distribution of the Frequency of Examiners' Team Mean Deviation Scores in Regard to Total Score in 2003

The total team deviation scores for each of the 9 teams are presented in Figure 25. Whereas Figure 24 was a summary of the number of examiners who were on teams that had a particular team mean deviation score, Figure 25 is a summary of how far each unique team's mean deviation score was from that team's total consensus score. Again, the closer the team mean deviation score was to zero, the more consistent that team's examiners' scores were considered to be. By observing the bars in the histogram, one can easily ascertain that examiners did not score consistently with the consensus score except for teams 22 and 25.

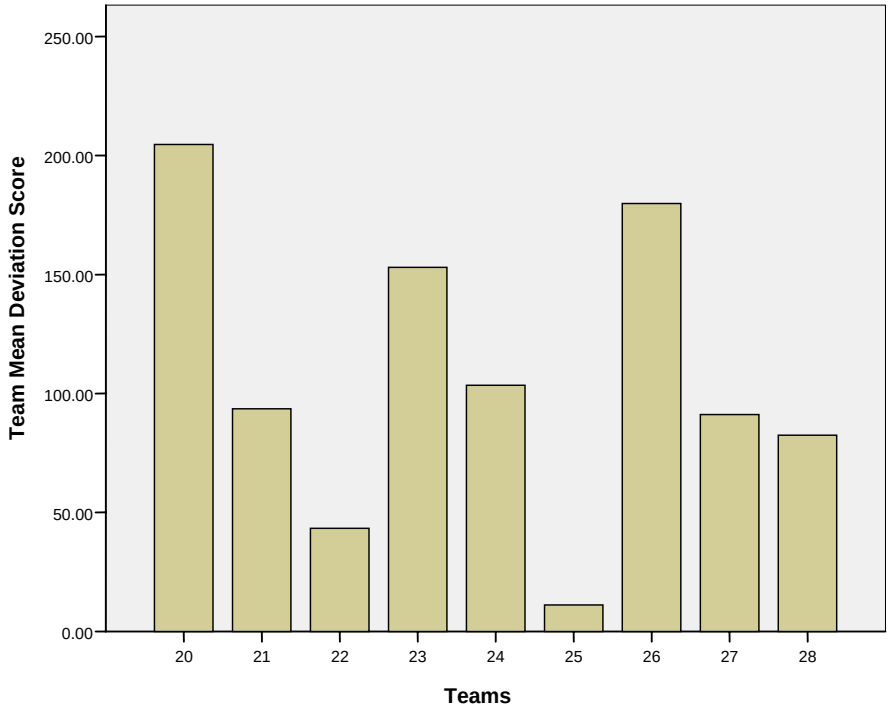

FIGURE 25. Distribution of Team Mean Deviation Scores by Teams in 2003

Team 22 was a new-level team (greater that 50% new examiners) that assessed a service organization. Team 25 was a senior-level team (greater than 50% returning and senior examiners) that also assessed a service organization. There does appear to be a pattern in that both teams assessed a service organization. Teams 20, 23, and 26 appear to share a commonality in that all three teams have a wide range of variation from their teams' respective total consensus scores. However, there is no pattern in sector or team experience level. One commonality does appear between teams 21 and 27. Both teams are new teams and both teams assessed health care organizations. Team 21 had a team mean deviation score of 93.57 and Team 27 had a team mean

deviation score of 91.9. Although the variation is large in comparison to the total consensus score, it is interesting that the two teams appear to be consistent in their variation. These were the only two teams in 2003 that assessed health care organizations. Possible causes and/or implications of these results will be further analyzed in Research Questions 2 and 3.

Descriptive statistics for 2004. There were 47 examiners in 6 teams for 2004. Considering that the closer the team mean deviation score is to zero, the more consistent examiners are considered to be, then scores above zero can be translated as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be translated as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

The team mean deviation was 97.70 with a standard deviation of 44.30. The skewness was -.692. In general, examiners did not produce individual total scores that were consistent with their team's total consensus scores. A summary of the descriptive statistics is presented in Table 14.

TABLE 14. Summary Statistics for Team Mean Deviation Scores for the Texas Award for Performance Excellence in 2004

Description	Statistics
Mean	97.7021
Variance	1962.178
Std. Deviation	44.29647
Minimum of	15.21
Maximum of	154.75
Skewness	-.692
Kurtosis	-.538

The 47 examiners were grouped into 6 teams. Each examiner had an individual total score which was used to generate an individual total deviation score from their team's total consensus score. These individual total deviation scores were then averaged to obtain a team mean deviation score for teach team. Figure 26 is a summary of the number of examiners who were on teams that had a particular team mean deviation score. The team mean deviation for the distribution in 2004 was 97.70, which indicates that examiners in 2004 scored more leniently when working alone than when coming to consensus as a team. The standard deviation was 44.30 and the range fell across a minimum of 15.21 and a maximum of 154.75 which indicated a large variation. All of the team mean deviations fell in a range above zero which makes 2004 data similar to 2003 and indicates that all examiners in all teams scored more leniently than the team consensus score.

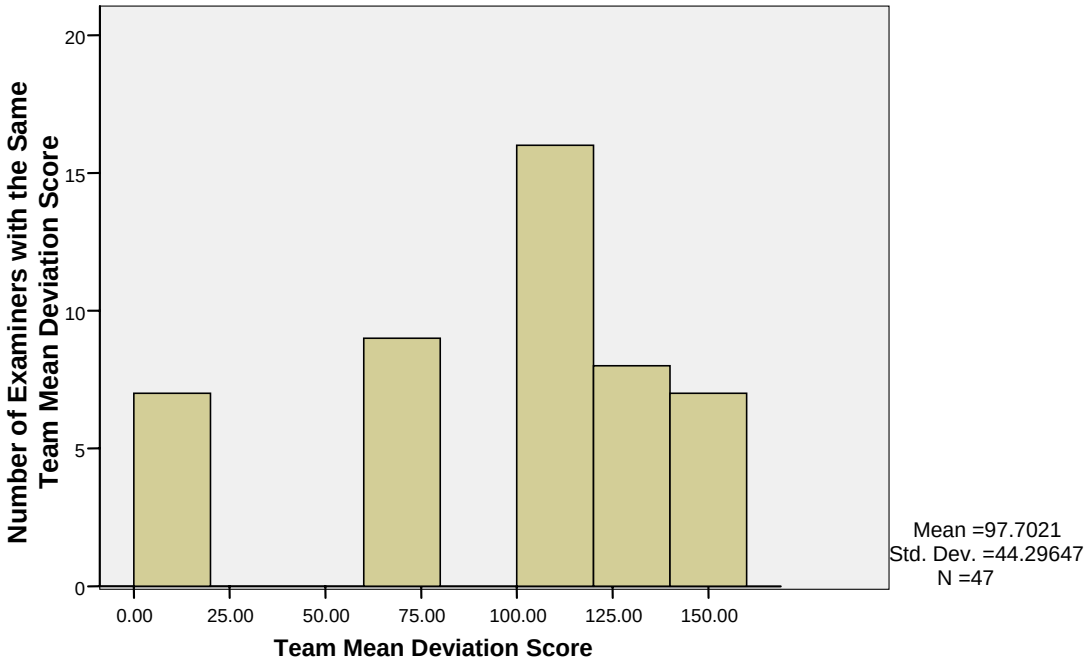

FIGURE 26. Distribution of the Frequency of Examiners' Team Mean Deviation Scores in Regard to Total Score in 2004

The total team deviation scores for each of the 6 teams are presented in Figure 27. Whereas Figure 26 was a summary of the number of examiners who were on teams that had a particular team mean deviation score, Figure 27 is a summary of how far each unique team's mean deviation score was from that team's total consensus score. Again, the closer the team mean deviation score was to zero, the more consistent that team's examiners' scores were considered to be. By observing the bars in the histogram, one can easily ascertain that examiners did not score consistently with the consensus score except for team 32. Team 32 was a senior-level team (greater than 50% returning and senior examiners) that assessed a health care organization. It

appears that teams 29, 31 and 33 have similar variation. Although all three teams were senior-level teams, they each assessed organizations from different sectors. Team 29 assessed an educational organization, team 31 assessed a public organization and team 33 assessed a manufacturing organization. Consistency in this case is not necessarily good because the variation from the total consensus score is so large.

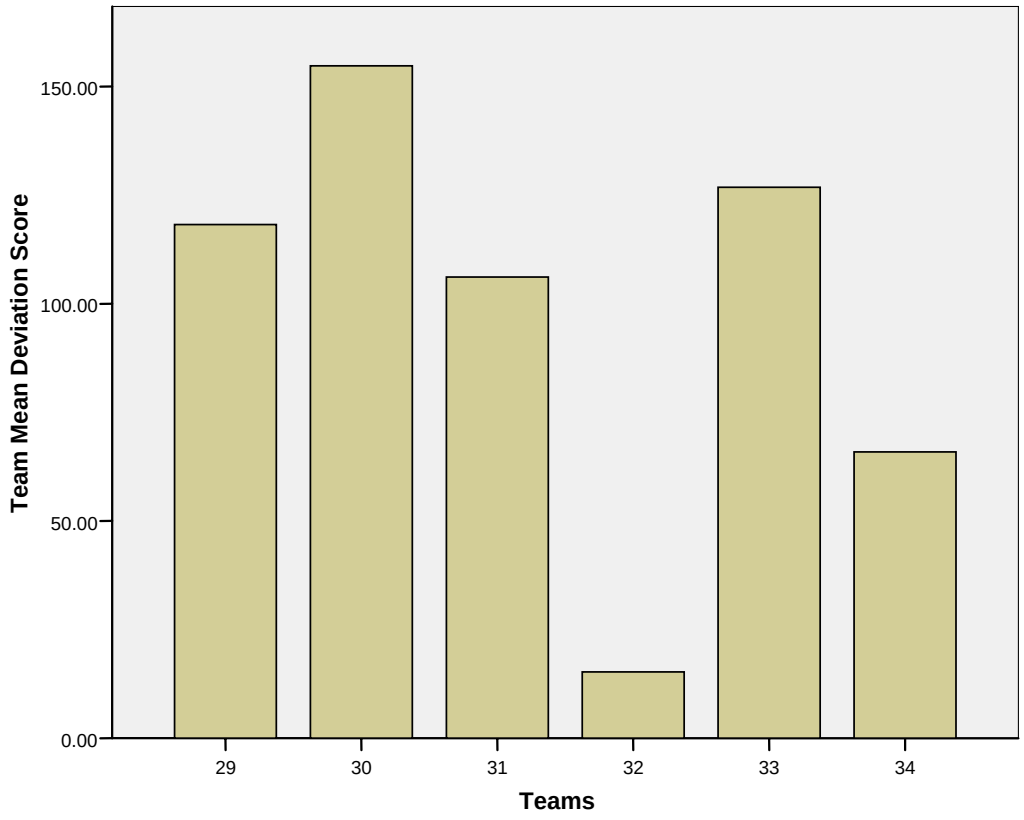

FIGURE 27. Distribution of Team Mean Deviation Scores by Teams in 2004

A summary of the means, standard deviations and skewness values is presented in Table 15. It appears that with each passing year, teams are scoring more consistently with the total consensus scores.

TABLE 15. Summary Table of Descriptive Statistics for the Texas Award for Performance Excellence for 2001-2004

Year	Mean	Std. Deviation	Skewness
2001	85.4474	58.69183	-.385
2002	106.9906	67.19948	-.446
2003	103.8651	59.24787	.160
2004	97.7021	44.29647	-.692

Summary of Research Question 1

Research Question 1 focused on the overall differences between individual total scores and team total consensus scores. The aggregated data and the data separated by each of the four years reveals that, in general, individual examiners did not produce scores that were consistent with their team's total consensus scores, and that, in general, individuals tended to score more leniently when working on their own than when coming to consensus as a team.

Only one team our of 34 had an overall team mean deviation score that was lower than the total consensus score as indicated in Figure 2. Furthermore, only 3 teams had team mean deviations scores that were close to the consensus score. All three teams were senior-level teams in terms of overall experience, but each team assessed organizations from different sectors and different years. Therefore, it would appear that the only influencing factor in consistency of scoring was that of team experience level. However, upon further analysis, it was discovered that three out of the four

teams with the greatest variation as indicated by team mean deviation scores were also senior-level teams. Again, each of the teams assessed organizations from different sectors and in different years. Analyses from each of the four years in isolation yielded similar results; examiners gave higher scores when assessing organizations independently than when coming to consensus as a team. Consequently, it was not possible to determine any influencing factors in consistency of examiner scoring based on results from Research Question 1 and the answer to the question of do the mean of the deviations of individual total scores from team total consensus scores equaled to zero was no.

One pattern that did emerge was that as the years progressed, the team mean deviation scores were growing smaller. This pattern may suggest several things. As understanding of the TAPE and the elements of continuous improvement are better understood, organizations themselves may be doing a better job of implementing continuous improvement. The pattern may also suggest that organizations may be doing a better job of completing the application making it possible for examiners to assess more consistently. Finally, the pattern may suggest that examiner training may be improving. More effective training of examiners could be what is leading to more consistency in scoring and resulting in scores being more consistent with the total consensus score. It is this possibility that is the focus of this study. Given the fact, however, that the data set is limited to only four years, it is difficult to make a determination. More data would be needed to establish whether or not the information in Table 15 is truly a trend.

Research Question 2 - Is the mean deviation between individual item scores and team item consensus scores equal to zero?

Research Question 2 was addressed through the development and analysis of descriptive statistics, histograms and other tables. Since team item consensus scores were considered the true score for each item, the closer the item mean deviation score was to zero, the more consistent in scoring the examiners were considered to be. For the purpose of this question, the mean of the item deviation scores was called the item mean deviation score. In order to determine the item mean deviation score, the team consensus item score was subtracted from each of the individual item scores to get an item deviation score for each item for each examiner. Next, the mean of the item deviation scores was calculated for each of the 17 items. An analysis was run for all 17 items across the 34 teams.

The summary statistics and histograms in Research Question 2 show one mean score for the item mean deviation scores. In some situations, it is not recommended to generate means of means (i.e. the mean of the item mean deviation scores); however, it was possible to do so in this case because all of the teams were relatively close in size. As previously noted, the teams ranged in size from 7 to 9 examiners. If there had been a wide range of team size, using the mean of a mean would not have been accurate or appropriate.

The number of items occasionally varied from year to year as the TAPE staff continually sought to improve the judging criterion and therefore made minor changes to the criteria when appropriate. Consequently, there were three items within the four years of collected data that were not consistent across 2001 through 2004.

Items 6.3, 7.5 and 7.6 were removed from the set of sample statistics leaving a total of 17 items in the data set. As stated previously, the alpha for the truncated data set was .959.

For this study, the closer the item mean deviation score was to zero, the more consistent in scoring examiners were considered to be. Therefore, scores above zero can be interpreted as examiners being more lenient when they score individually than when they come to consensus with a team. Scores below zero can be interpreted as examiners being more stringent or critical when they score individually than when they come to consensus with a team.

Item 1.1 organizational leadership. Table 16 is a summary of the descriptive statistics for Item 1.1 – Organizational Leadership. An item mean deviation score for Item 1.1 was calculated for each of the 34 teams.

TABLE 16. Summary Statistics for Item Mean Deviation Scores for Item 1.1 - Organizational Leadership

Description	Statistics
Mean	-6.8064
Std. Deviation	16.67523
Minimum of	-74.00
Maximum of	40.00
Skewness	-.427
Kurtosis	.675

The mean for all the teams was -6.81 and the standard deviation was 16.68, which indicates that examiners tended to score a little more critically when evaluating

organizational leadership on their own than when working with a team to come to consensus. There was a great deal of variation, however, as evidenced by the range that fell across a minimum of -74 and maximum of 40.

Figure 28 is a summary of frequency with which the 250 examiners produced the same item mean deviation score for Item 1.1. The item mean deviation score for the distribution for Item 1.1 was -6.81, which indicates that examiners scored slightly more stringently when working alone than when coming to consensus as a team. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 1.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Further investigation revealed approximately 86% of all scores were within +/- 20 points of zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores seem to be less stable, as evidenced by the variation of the length of the bars on the histogram.

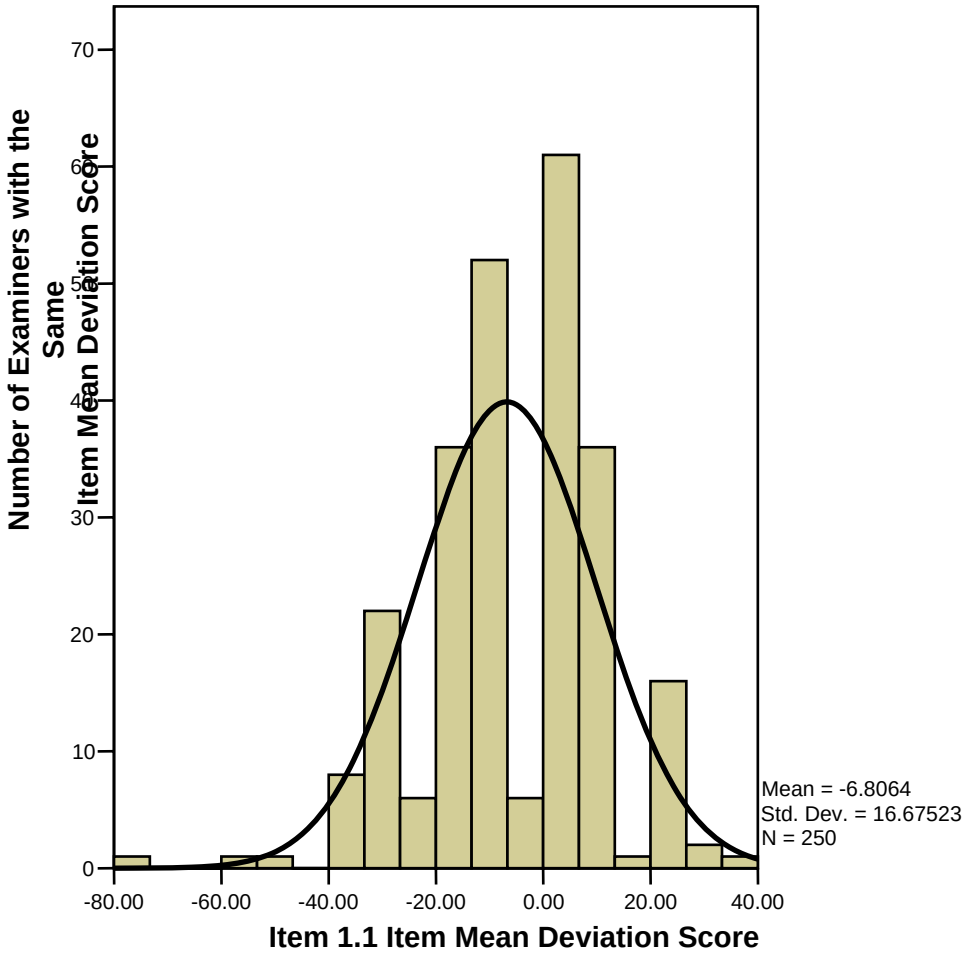

FIGURE 28. Frequency Distribution of Item Mean Deviation Scores for Item 1.1 – Organizational Leadership

Item 1.2 social responsibility. Table 17 is a summary of the descriptive statistics for Item 1.2 – Social Responsibility. An item mean deviation score for Item 1.2 was calculated for each of the 34 teams.

The mean for all the teams was -7.57 and the standard deviation was 17.89, which indicates that examiners tended to score a little more critically when evaluating social responsibility on their own than when working with a team to come to consensus.

However, there was a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 40.

TABLE 17. Summary Statistics for Item Mean Deviation Scores for Item 1.2 – Social Responsibility

Description	Statistics
Mean	-7.5712
Std. Deviation	17.89411
Minimum of	-70.00
Maximum of	40.00
Skewness	-.421
Kurtosis	.339

Figure 29 is a summary of frequency with which the 250 produced the same item mean deviation score for Item 1.2. The item mean deviation score for the distribution for Item 1.2 was -7.57, which indicates that examiners scored slightly more stringently when working alone than when coming to consensus as a team. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 1.2, the most prevalent item mean deviation scores were within +/- 10 points of zero. Further investigation revealed approximately 85% of the item mean deviation scores were within +/- 20 points of zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores seem to be less stable, as evidenced by the variation of the length of the bars on the histogram.

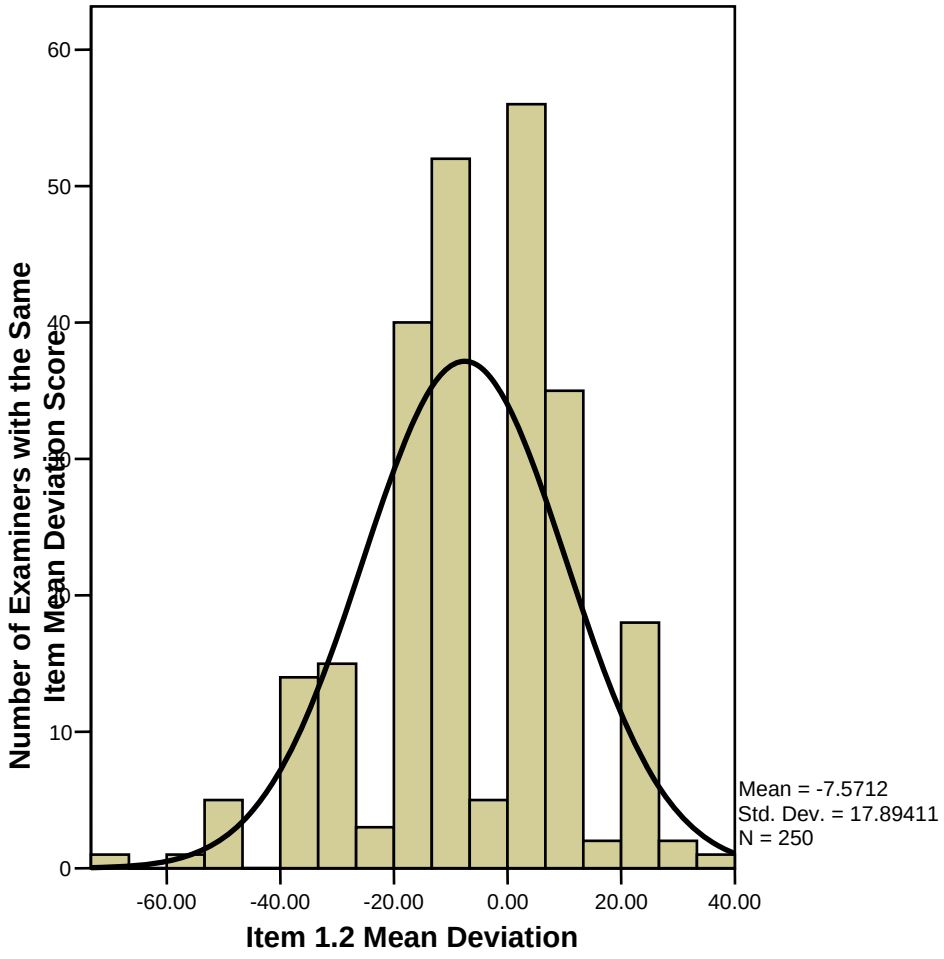

FIGURE 29. Frequency Distribution of Item Mean Deviation Scores for Item 1.2 – Social Responsibility

Item 1.1 and 1.2 were the two items that made up the Leadership category. Although there was variation within each item, examiners did score relatively consistently across both. Item 1.1 and Item 1.2 were assessed with a consistent approach, however highlighted results for item 1.2 expressed across a tighter range of +/- 10 illustrate the item's smaller variation in comparison to Item 1.1. The similar means and skewness values indicate that examiners assessed leadership with a similar approach.

Item 2.1 strategy development. Table 18 is a summary of the descriptive statistics for Item 2.1 – Strategy Development. An item mean deviation score for Item 2.1 was calculated for each of the 34 teams. The mean for all the teams was -9.67 and the standard deviation was 17.90, which indicates that examiners tended to score slightly more critically when evaluating strategy development on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 30.

TABLE 18. Summary Statistics for Item Mean Deviation Scores for Item 2.1 – Strategy Development

Description	Statistics
Mean	-9.6712
Std. Deviation	17.90451
Minimum of	-70.00
Maximum of	30.00
Skewness	-.577
Kurtosis	.515

Figure 30 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 2.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 83% of all scores were within +/- 20 points of zero. Therefore, although there was variation overall, the majority of examiners assessed

the item closer to the team item consensus score than those who did not. As the scores move farther away from zero and more to the critical side, the scores seem to be less stable, as evidenced by the variation in length of the bars on the histogram.

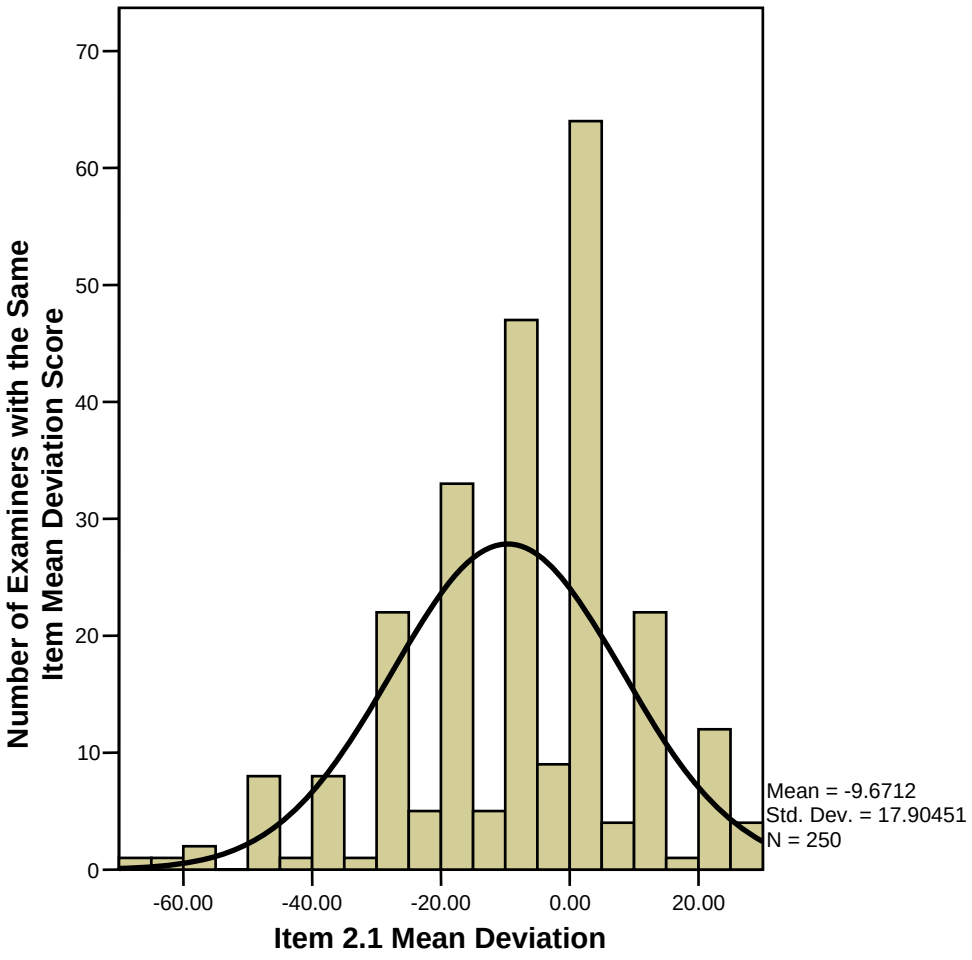

FIGURE 30. Frequency Distribution of Item Mean Deviation Scores for Item 2.1 - Strategy Development

Item 2.2 strategy deployment. Table 19 is a summary of the descriptive statistics for Item 2.2 – Strategy Deployment. An item mean deviation score for Item 2.2 was

calculated for each of the 34 teams. The mean for all the teams was -9.45 and the standard deviation was 18.10, which indicates that examiners tended to score slightly more critically when evaluating strategy deployment on their own than when working with a team to come to consensus. However, there was a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 20.

TABLE 19. Summary Statistics for Item Mean Deviation Scores for Item 2.2 – Strategy Deployment

Description	Statistics
Mean	-9.4504
Std. Deviation	18.05830
Minimum of	-80.00
Maximum of	20.00
Skewness	-.779
Kurtosis	.687

Figure 31 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores is most prevalent. For Item 2.2, the most prevalent item team mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 83 % of the item mean deviation fell between +/- 20 points of zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores seem to be less stable, as evidenced by the variation in length of the bars on the histogram.

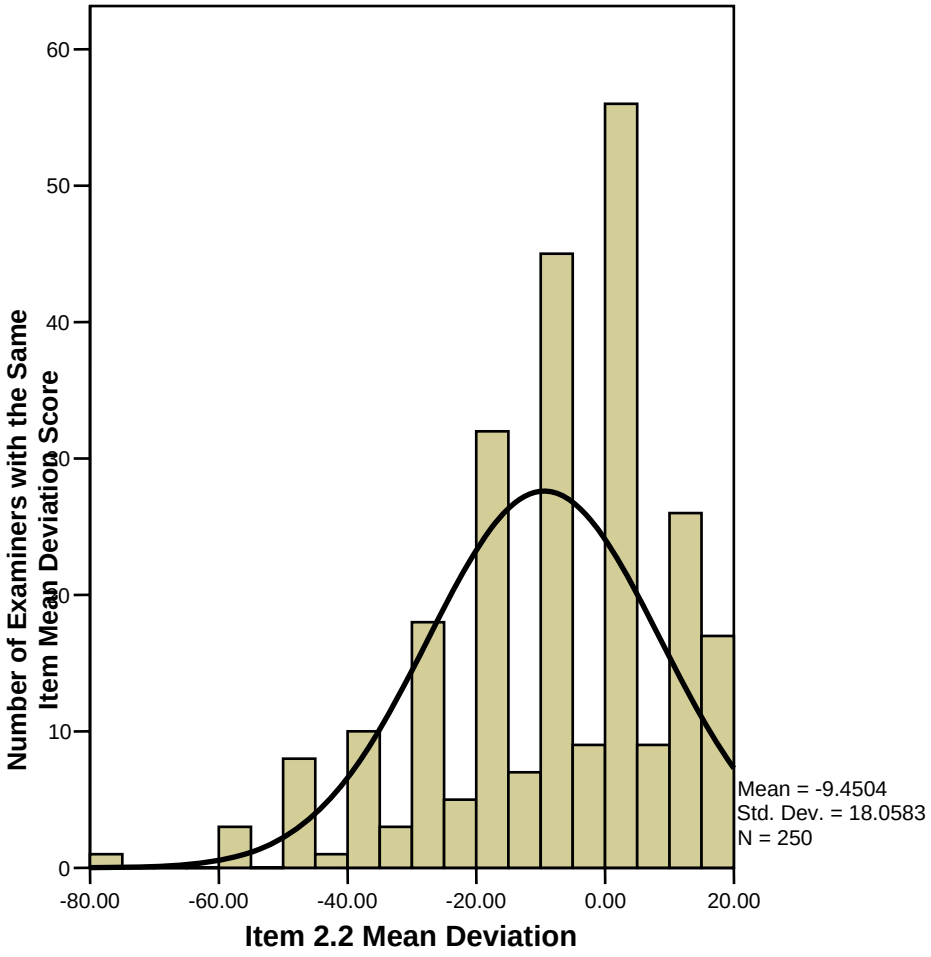

FIGURE 31. Frequency Distribution for Item Mean Deviation Scores for Item 2.2 – Strategy Deployment

Items 2.1 and 2.2 were the two items that made up the Strategic Planning category. Although there was variation within each item, examiners did score relatively consistently across both. Examiner assessments of both items resulted in 83% of the item mean deviation scores falling within 20 points of zero. The similar means and skewness values also indicate that examiners assessed strategic planning with a similar approach.

Item 3.1 customer and market knowledge. Table 20 is a summary of the descriptive statistics for Item 3.1 – Customer and Market Knowledge. An item mean deviation score for Item 3.1 was calculated for each of the 34 teams. The mean for all the teams was -9.52 and the standard deviation was 17.88, which indicates that examiners tended to score slightly more critically when evaluating customer and market knowledge on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 30.

TABLE 20. Summary Statistics for Item Mean Deviation Scores for Item 3.1 – Customer and Market Knowledge

Description	Statistics
Mean	-9.5248
Std. Deviation	17.88286
Minimum of	-80.00
Maximum of	30.00
Skewness	-.694
Kurtosis	.994

Figure 32 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores was most prevalent. For Item 3.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that 84% of the item mean deviation scores were within +/- 20 points of zero. Therefore, although there was variation overall, the

majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores seem to be less stable, as evidenced by the variation in length of the bars on the histogram.

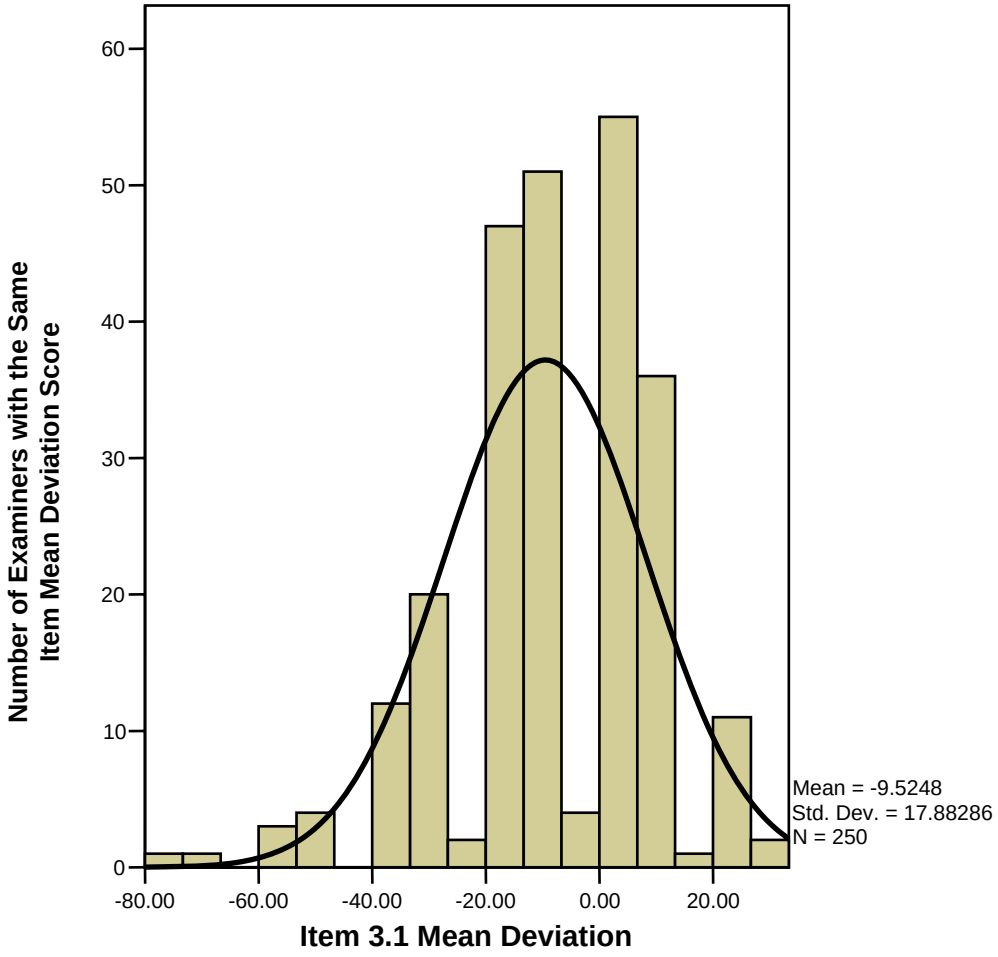

FIGURE 32. Frequency Distribution of Item Mean Deviation Scores for Item 3.1 – Customer and Market Knowledge

Item 3.2 customer relationship and satisfaction. Table 21 is a summary of the descriptive statistics for Item 3.2 – Customer Relationship and Satisfaction. An item mean deviation score for Item 3.1 was calculated for each of the 34 teams. The mean

for all the teams was -8.30 and the standard deviation was 16.61, which indicates that examiners tended to score slightly more critically when evaluating customer relationship and satisfaction on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -77 and a maximum of 36.25.

TABLE 21. Summary Statistics for Mean Deviation Scores for Item 3.2 – Customer Relationship and Satisfaction

Description	Statistics
Mean	-8.3024
Std. Deviation	16.61119
Minimum of	-77.00
Maximum of	36.25
Skewness	-.663
Kurtosis	1.223

Figure 33 is a summary of frequency with which 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 3.2, the most prevalent item mean deviation scores were within +/- 20 points of zero.

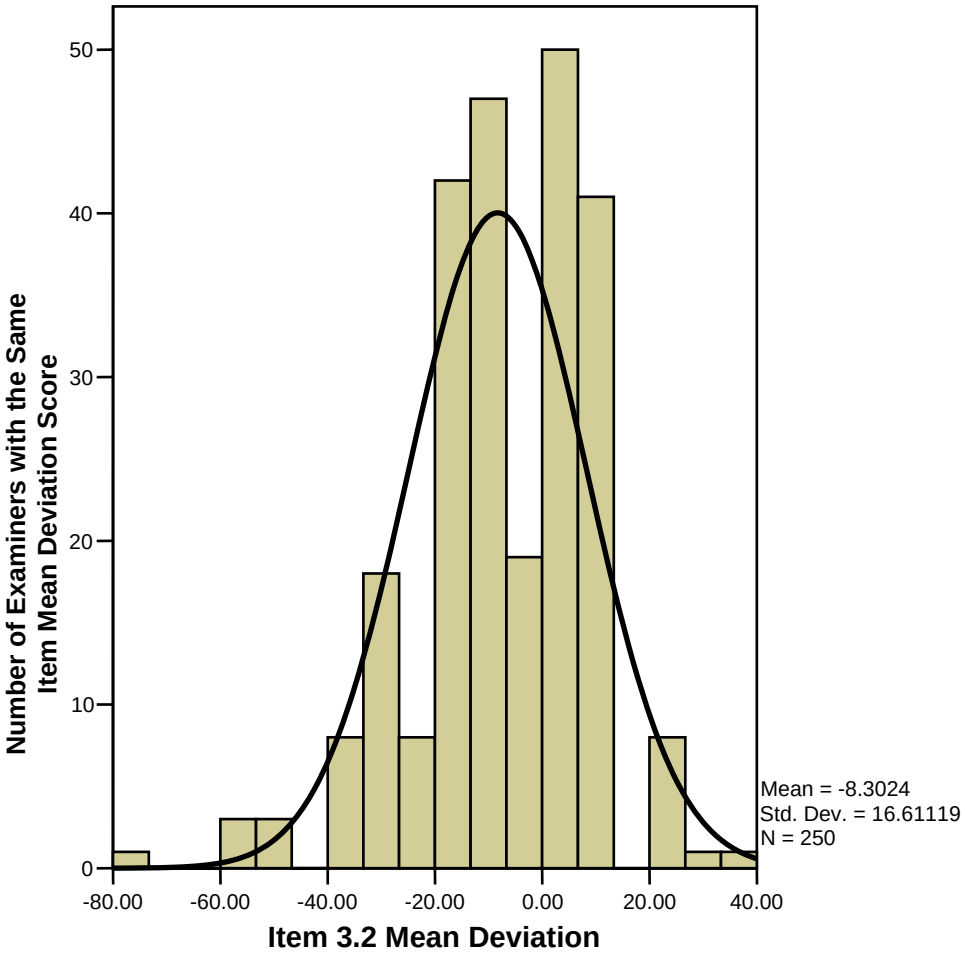

FIGURE 33. Frequency Distribution for Item Mean Deviation Scores for Item 3.2 – Customer Relationship and Satisfaction

Upon further investigation, it was revealed that 86% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

Items 3.1 and 3.2 were the two items that made up the Customer and Market Focus category. Although there was variation within each item, examiners did score

relatively consistently across both. Examiner assessments of both items resulted more than 80% of the scores falling within +/- 20 points of zero. The similar means and skewness values also indicated that examiners assessed customer and market focus with a similar approach.

Item 4.1 measurement and analysis of organizational performance. Table 22 is a summary of the descriptive statistics for Item 4.1 – Measurement and Analysis of Organizational Performance. An item mean deviation score for Item 4.1 was calculated for each of the 34 teams. The mean for all the teams was -6.83 and the standard deviation was 17.03, which indicates that examiners tended to score slightly more critically when evaluating measurement and analysis of organizational performance on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -63 and a maximum of 30.

Figure 34 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 4.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 85% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

TABLE 22. Summary Statistics for Item Mean Deviation Scores for Item 4.1 - Measurement and Analysis of Organizational Performance

Description	Statistics
Mean	-6.8320
Std. Deviation	17.03722
Minimum of	-63.00
Maximum of	30.00
Skewness	-.490
Kurtosis	.526

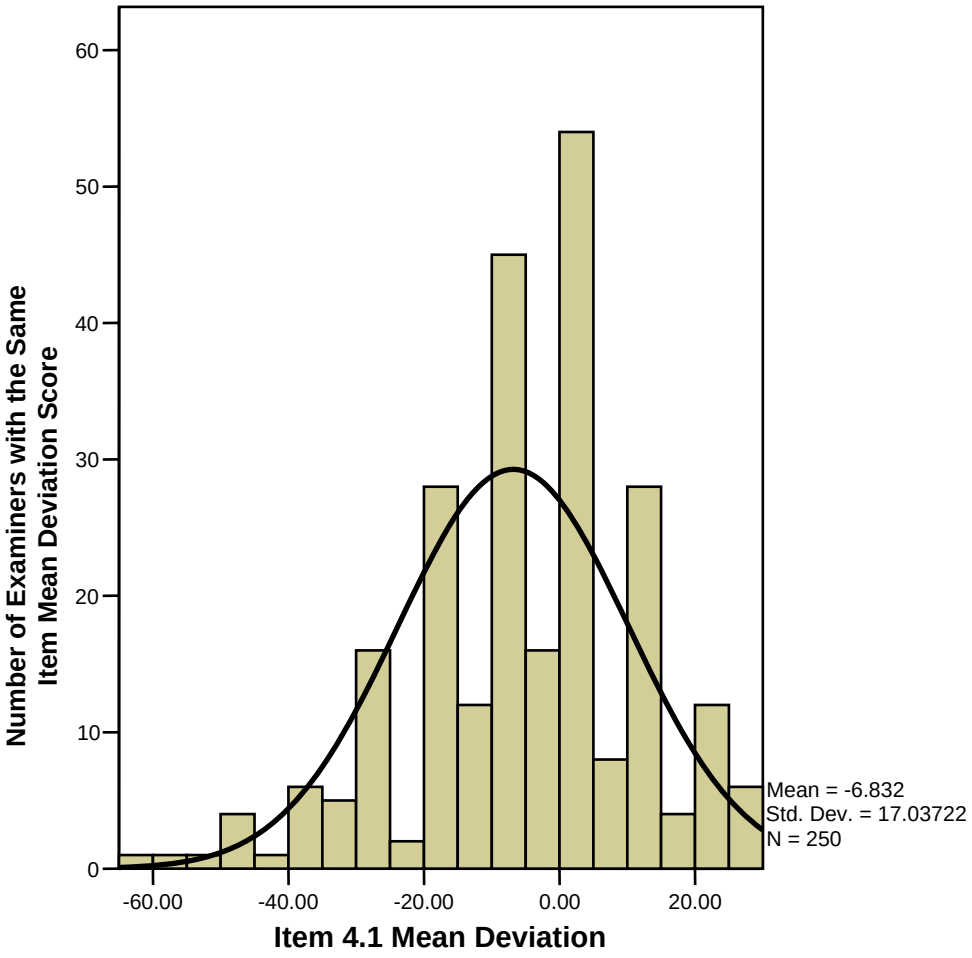

FIGURE 34. Frequency Distribution for Item Mean Deviation Scores for Item 4.1 – Measurement and Analysis of Organizational Performance

Item 4.2 information and knowledge management. Table 23 is a summary of the descriptive statistics for Item 4.2 – Information and Knowledge Management. An item mean deviation score for Item 4.2 was calculated for each of the 34 teams. The mean for all the teams was -8.69 and the standard deviation was 17.73, which indicates that examiners tended to score slightly more critically when evaluating information and knowledge management on their own than when working with a team to come to consensus. However, there was a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 30.

TABLE 23. Summary Statistics for Item Mean Deviation Scores for Item 4.2 – Information and Knowledge Management

Description	Statistics
Mean	-8.6936
Std. Deviation	17.72560
Minimum of	-70.00
Maximum of	30.00
Skewness	-.616
Kurtosis	.553

Figure 35 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 4.2, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 83% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those

who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

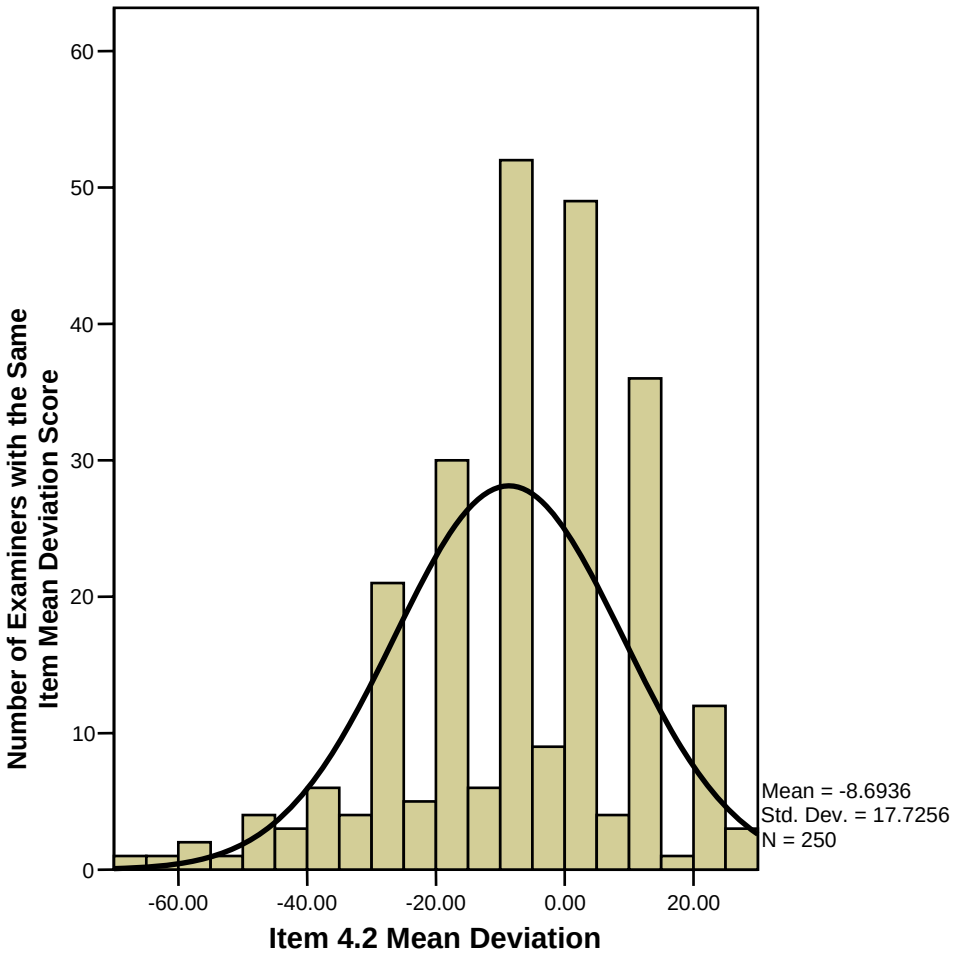

FIGURE 35. Frequency Distribution for Item Mean Deviation Scores for Item 4.2 – Information and Knowledge Management

Items 4.1 and 4.2 were the two items that made up the Measurement, Analysis, and Knowledge Management category. Although there was variation within each item, examiners did score relatively consistently across both. Examiner assessments

of both items resulted more than 80% of the scores falling within +/- 20 points of zero. The similar means and skewness values indicate that examiners assessed measurement, analysis, and knowledge management with a similar approach.

Item 5.1 work systems. Table 24 is a summary of the descriptive statistics for Item 5.1 – Work Systems. An item mean deviation score for Item 5.1 was calculated for each of the 34 teams. The mean for all the teams was -6.13 and the standard deviation was 15.77, which indicates that examiners tended to score slightly more critically when evaluating work systems on their own than when working with a team to come to consensus. However, there was a great deal of variation as evidenced by the range that fell across a minimum of -60 and a maximum of 30.

TABLE 24. Summary Statistics for Item Mean Deviation Scores for Item 5.1 – Work Systems

Description	Statistics
Mean	-6.1344
Std. Deviation	15.76754
Minimum of	-60.00
Maximum of	30.00
Skewness	-.442
Kurtosis	.215

Figure 36 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 5.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 88% of the item mean deviation scores were within

+/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

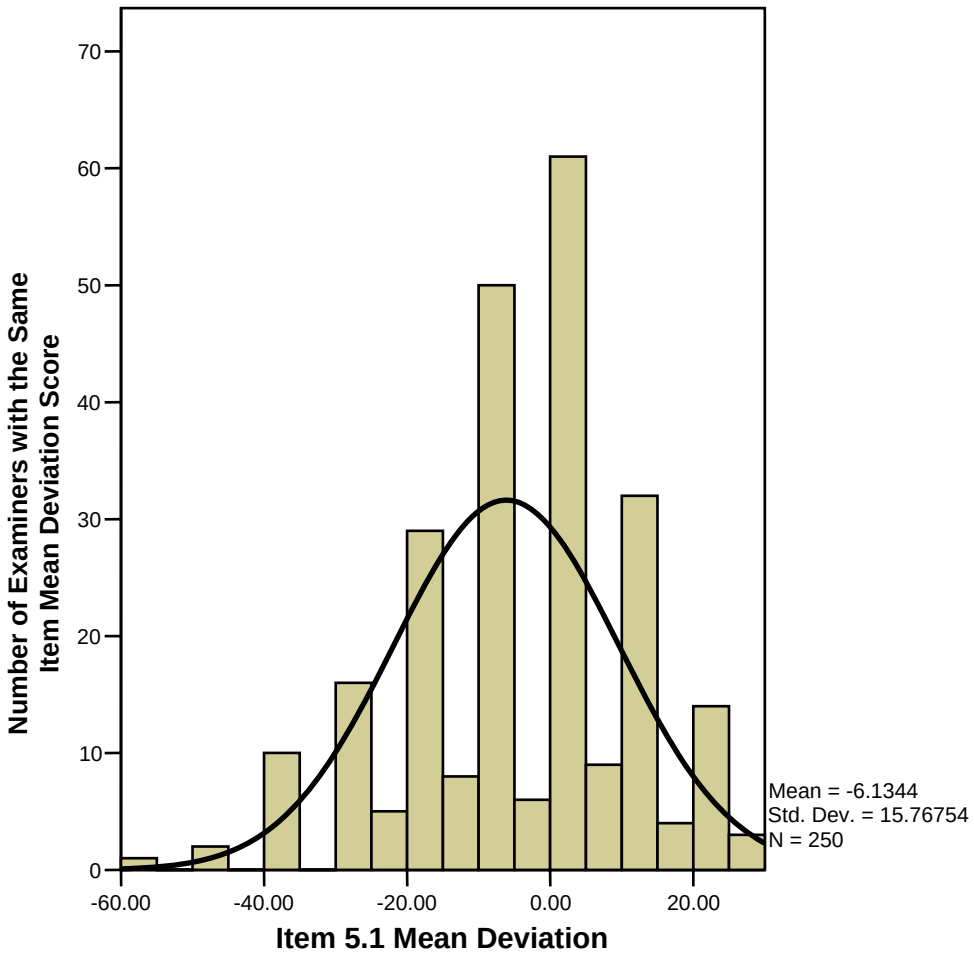

FIGURE 36. Frequency Distribution for Item Mean Deviation Scores for Item 5.1 – Work Systems

Item 5.2 employee learning & motivation. Table 25 is a summary of the descriptive statistics for Item 5.2 – Employee Learning & Motivation. An item mean

deviation score for Item 5.2 was calculated for each of the 34 teams. The mean for all the teams was -5.07 and the standard deviation was 16.07, which indicates that examiners tended to score slightly more critically when evaluating employee learning and motivation on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 30.

TABLE 25. Summary Statistics for Item Mean Deviation Scores for Item 5.2 – Employee Learning and Motivation

Description	Statistics
Mean	-5.0680
Std. Deviation	16.06961
Minimum of	-70.00
Maximum of	30.00
Skewness	-.802
Kurtosis	1.137

Figure 37 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 5.2, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 88% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

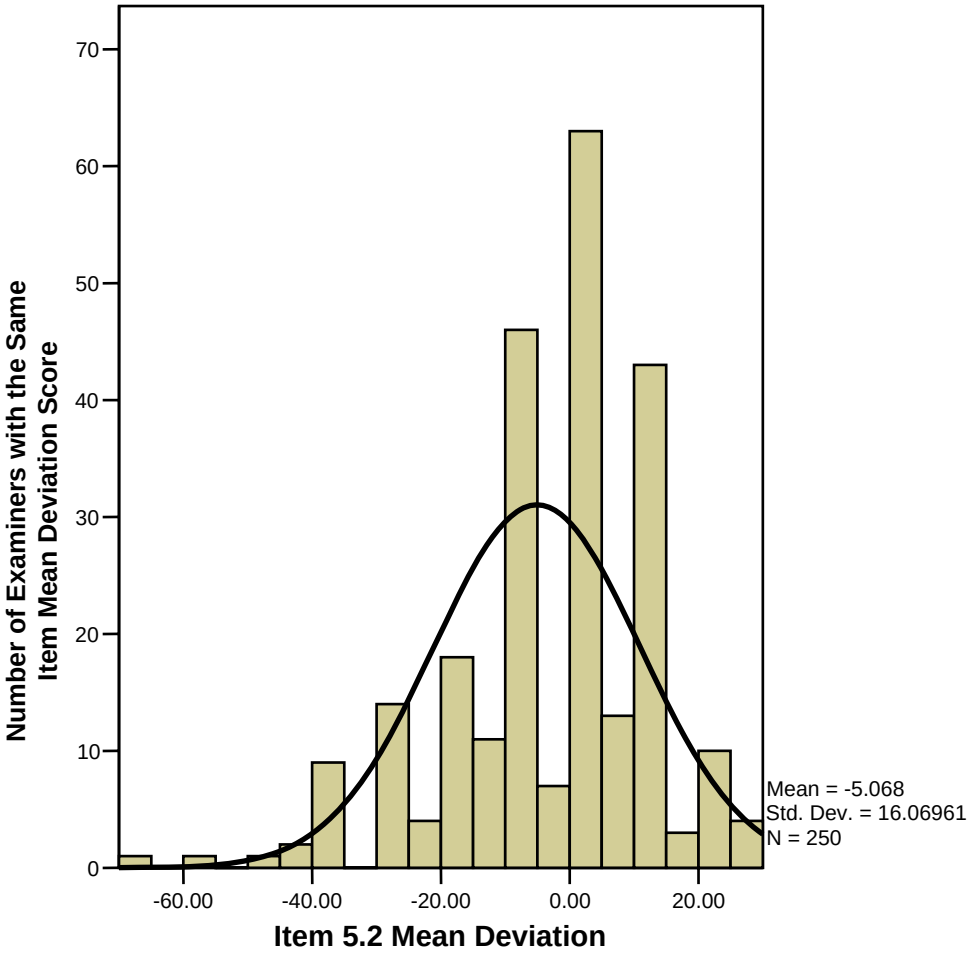

FIGURE 37. Frequency Distribution for Item Mean Deviation Scores for Item 5.2 – Employee Learning and Motivation

Item 5.3 employee well-being and satisfaction. Table 26 is a summary of the descriptive statistics for Item 5.3 – Employee Well-Being and Satisfaction. An item mean deviation score for Item 5.3 was calculated for each of the 34 teams. The mean for all the teams was -6.35 and the standard deviation was 17.11, which indicates that examiners tended to score slightly more critically when evaluating employee well-being and satisfaction on their own than when working with a team to come to

consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 30.

TABLE 26. Summary Statistics for Item Mean Deviation Scores for Item 5.3 – Employee Well-Being and Satisfaction

Description	Statistics
Mean	-6.3496
Std. Deviation	17.10928
Minimum of	-80.00
Maximum of	30.00
Skewness	-.827
Kurtosis	1.256

Figure 38 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 5.3, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 88% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

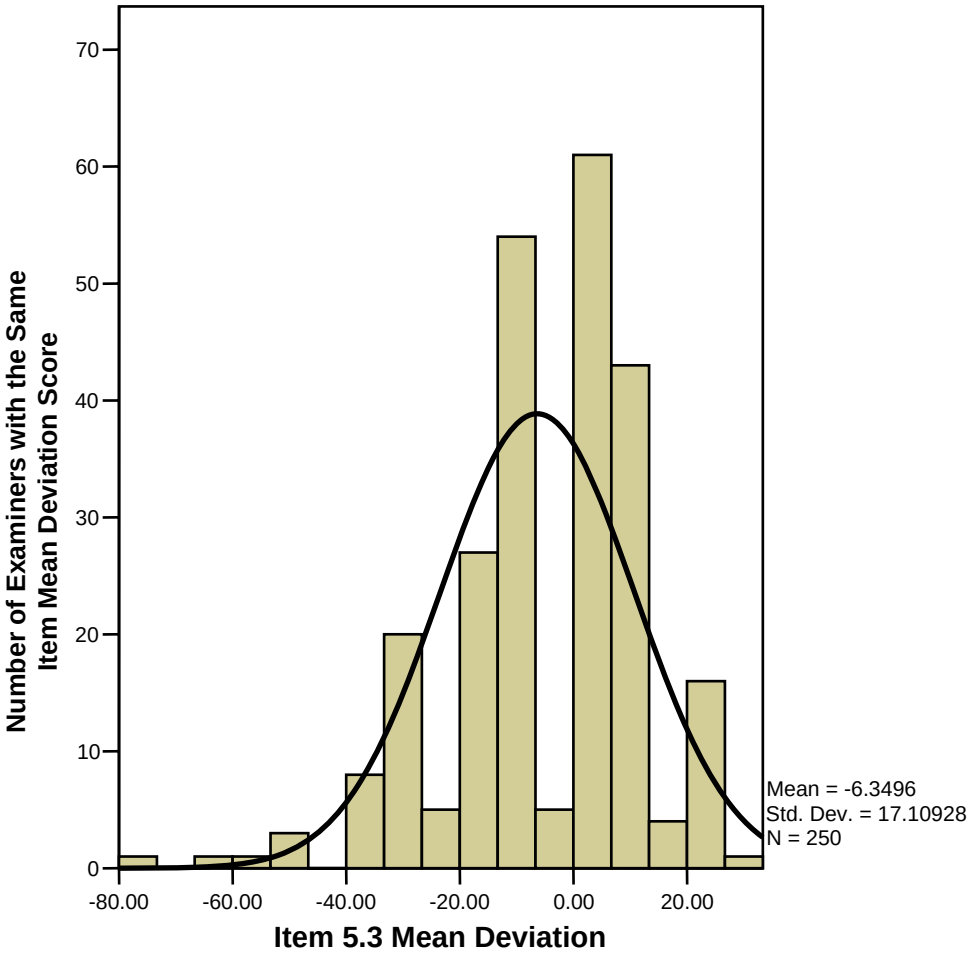

FIGURE 38. Frequency Distribution for Item Mean Deviation Scores for Item 5.3 – Employee Well-being and Satisfaction

Items 5.1, 5.2 and 5.3 were the three items that made up the Human Resource Focus category. Although there was variation within each item, examiners did score relatively consistently across all three. Examiner assessments resulted in 88% of the item mean deviation scores falling within +/- 20 points of zero for each of the three items. The similar means and skewness values indicate that examiners assessed human resource focus with a similar approach.

Item 6.1 value creation processes. Table 27 is a summary of the descriptive statistics for Item 6.1 – Value Creation Processes. An item mean deviation score for Item 6.1 was calculated for each of the 34 teams. The mean for all the teams was -8.43 and the standard deviation was 17.66, which indicates that examiners tended to score slightly more critically when evaluating value creation processes on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 37.50.

TABLE 27. Summary Statistics for Item Mean Deviation Scores for Item 6.1 – Value Creation Processes

Description	Statistics
Mean	-8.4288
Std. Deviation	17.65580
Minimum of	-70.00
Maximum of	37.50
Skewness	-.553
Kurtosis	.565

Figure 39 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 6.1, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 86% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those

who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

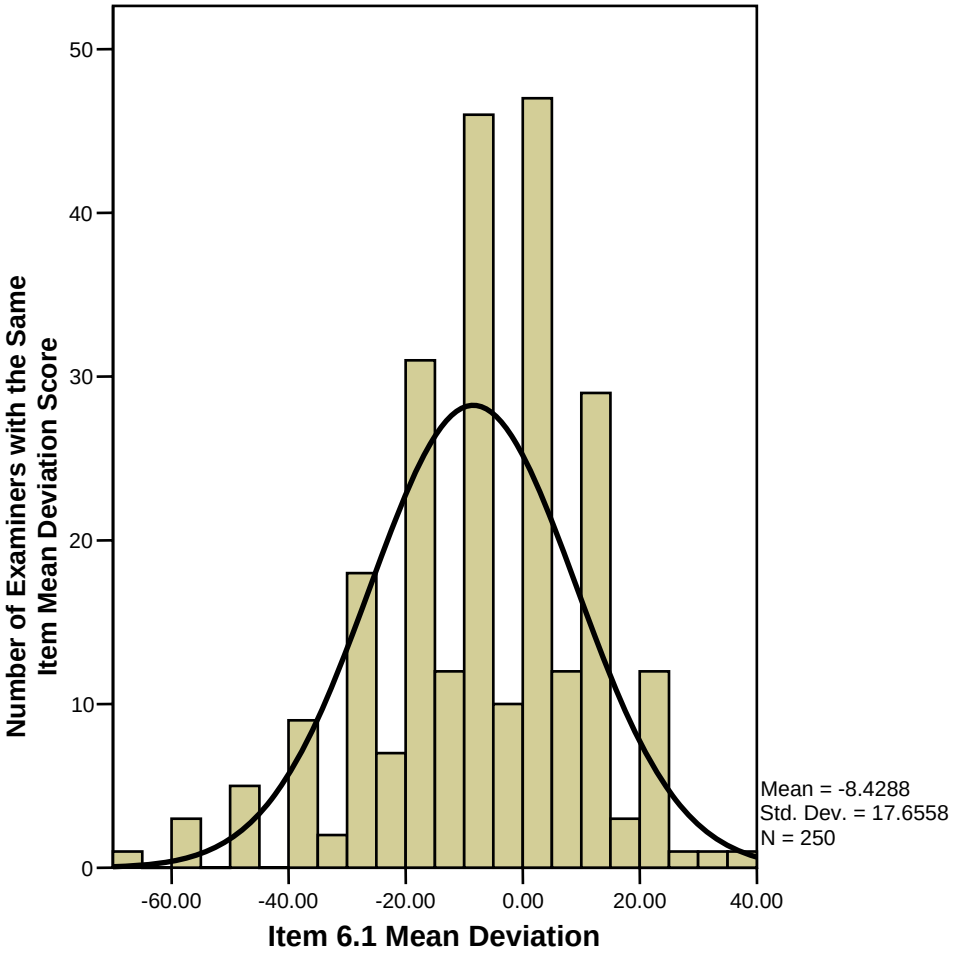

FIGURE 39. Frequency Distribution for Item Mean Deviation Scores for Item 6.1 – Value Creation Processes

Item 6.2 support processes. Table 28 is a summary of the descriptive statistics for Item 6.2 – Support Processes. An item mean deviation score for Item 6.2 was calculated for each of the 34 teams. The mean for all the teams was -10.02 and the standard deviation was 19.06, which indicates that examiners tended to score slightly

more critically when evaluating Support Processes on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -70 and a maximum of 30.

TABLE 28. Summary Statistics for Item Mean Deviation Scores for Item 6.2 – Support Processes

Description	Statistics
Mean	-10.0240
Std. Deviation	19.06092
Minimum of	-70.00
Maximum of	30.00
Skewness	-.616
Kurtosis	.375

Figure 40 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 6.2, the most prevalent item mean deviation scores were within +/- 10 points of zero. Upon further investigation, it was revealed that approximately 63% of the item mean deviation scores were within +/- 10 points from zero, and 80% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

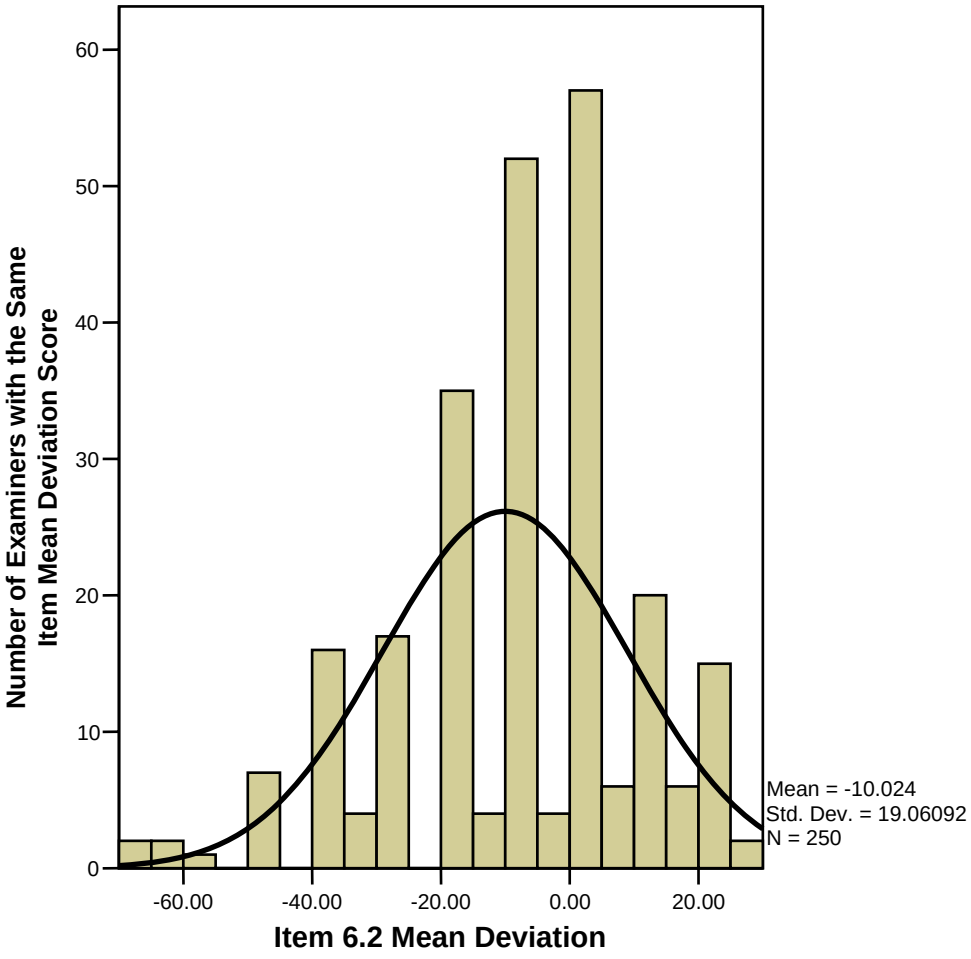

FIGURE 40. Frequency Distribution for Item Mean Deviation Scores for Item 6.2 – Support Processes

Items 6.1 and 6.2 were the three items that made up the Process Management category. Although there was variation within each item, examiners did score relatively consistently across both. Examiner assessments resulted in more than 80% of the item mean deviation scores falling within +/- 20 points of zero for both items. The similar means and skewness values indicate that examiners assessed human resource focus with a similar approach.

Item 7.1 customer-focused results. Table 29 is a summary of the descriptive statistics for Item 7.1 – Customer Focused Results. An item mean deviation score for Item 7.1 was calculated for each of the 34 teams. The mean for all the teams was -12.12 and the standard deviation was 17.25, which indicates that examiners tended to score slightly more critically when evaluating Customer Focused Results on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 20.

TABLE 29. Summary Statistics for Item Mean Deviation Scores for Item 7.1 – Customer-focused Results

Description	Statistics
Mean	-12.1240
Std. Deviation	17.25996
Minimum of	-80.00
Maximum of	20.00
Skewness	-.943
Kurtosis	1.480

Figure 41 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 7.1, the most prevalent item mean deviation scores were within +/- 10 points of zero. Upon further investigation, it was revealed that approximately 64% of the item mean deviation scores were within +/- 10 points from zero, and 80% of the item mean deviation scores were within +/-

20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

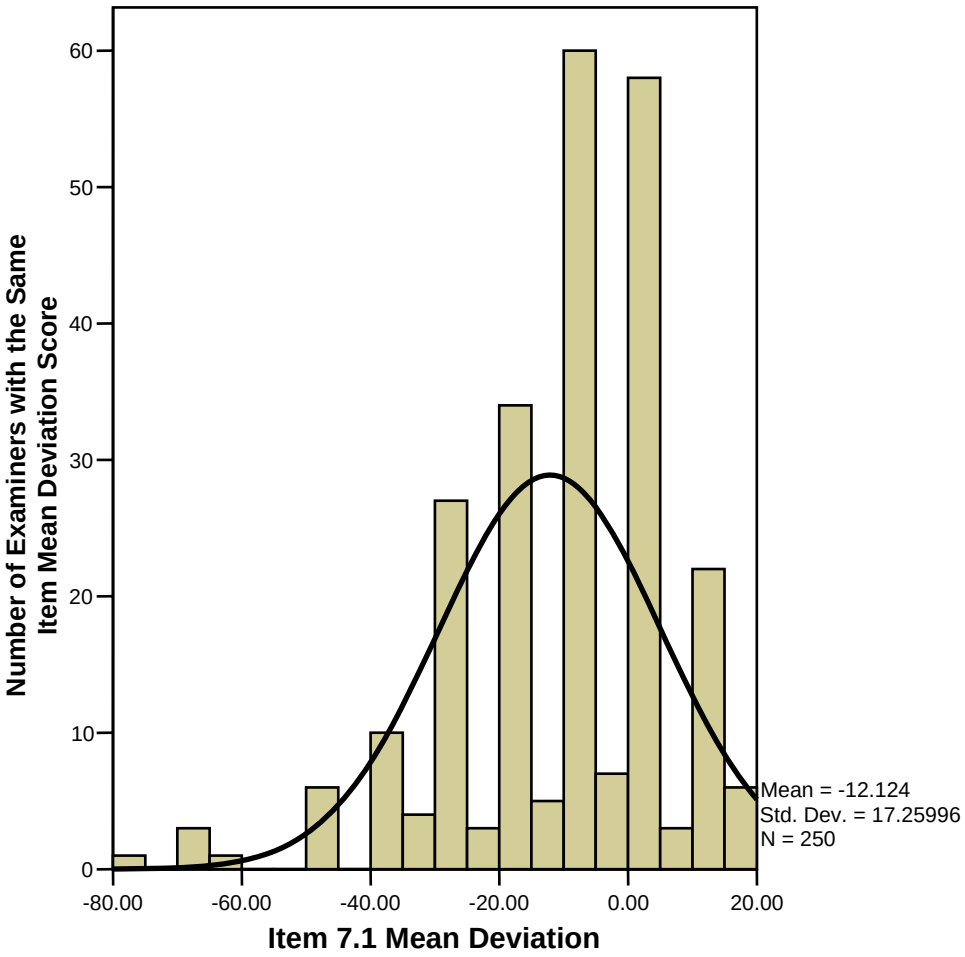

FIGURE 41. Frequency Distribution for Item Mean Deviation Scores for Item 7.1 – Customer-focused Results

Item 7.2 product and service results. Table 30 is a summary of the descriptive statistics for Item 7.2 – Product and Service Results. An item mean deviation score for Item 7.2 was calculated for each of the 34 teams. The mean for all the teams was -11.56 and the standard deviation was 17.42, which indicates that examiners tended to score slightly more critically when evaluating Product and Service Results on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -90 and a maximum of 30.

TABLE 30. Summary Statistics for Item Mean Deviation Scores for Item 7.2 – Product and Service Results

Description	Statistics
Mean	-11.5648
Std. Deviation	17.42318
Minimum of	-90.00
Maximum of	30.00
Skewness	-.748
Kurtosis	1.756

Figure 42 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For Item 7.2, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 79% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those

who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

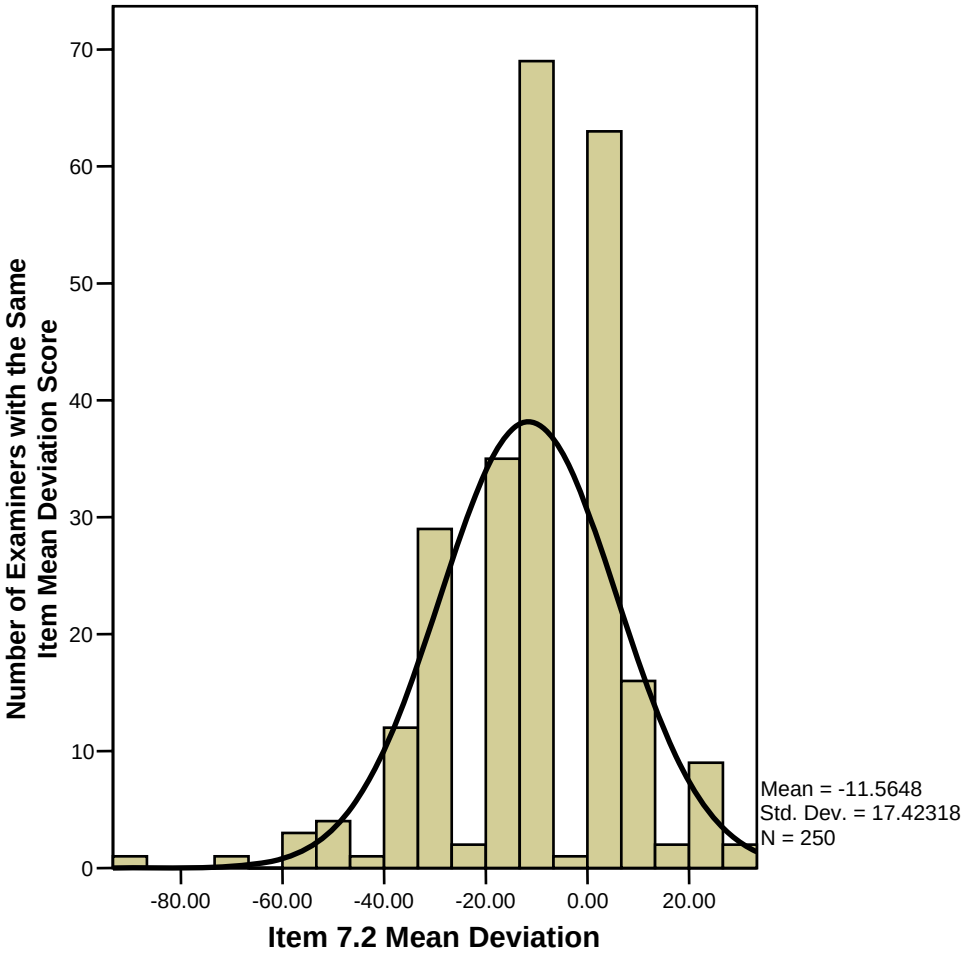

FIGURE 42. Frequency Distribution for Item Mean Deviation Scores for Item 7.2 – Product and Service Results

Item 7.3 financial and market results. Table 31 is a summary of the descriptive statistics for Item 7.3 Financial and Market Results. An item mean deviation score for Item 7.3 was calculated for each of the 34 teams. The mean for all the teams was -

11.53 and the standard deviation was 16.15, which indicates that examiners tended to score slightly more critically when evaluating Financial and Market Results on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 20.

TABLE 31. Summary Statistics for Item Mean Deviation Scores for Item 7.3 – Financial and Market Results

Description	Statistics
Mean	-11.5288
Std. Deviation	16.14555
Minimum of	-80.00
Maximum of	20.00
Skewness	-1.134
Kurtosis	2.317

Figure 43 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores. For item 7.3, the most prevalent item mean deviation scores were within +/- 20 points of zero.

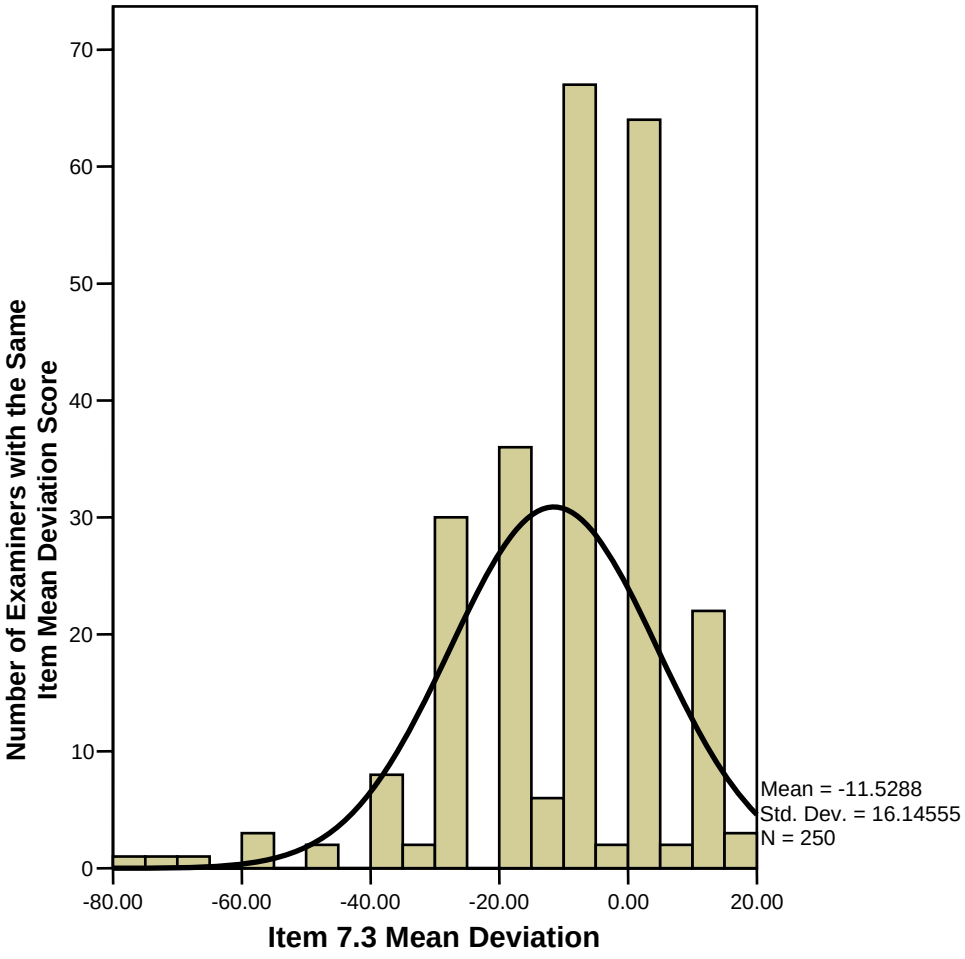

FIGURE 43. Frequency Distribution for Item Mean Deviation Scores for Item 7.3 – Financial and Market Results

Upon further investigation, it was revealed that approximately 82% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

Item 7.4 human resource results. Table 32 is a summary of the descriptive statistics for Item 7.4 Human Resource Results. An item mean deviation score for Item 7.4 was calculated for each of the 34 teams. The mean for all the teams was -12.49 and the standard deviation was 17.77, which indicates that examiners tended to score more critically when evaluating Human Resource Results on their own than when working with a team to come to consensus. There was, however, a great deal of variation as evidenced by the range that fell across a minimum of -80 and a maximum of 20.

TABLE 32. Summary Statistics for Item Mean Deviation Scores for Item 7.4 – Human Resource Results

Description	Statistics
Mean	-12.4864
Std. Deviation	17.76915
Minimum of	-80.00
Maximum of	20.00
Skewness	-.870
Kurtosis	1.123

Figure 44 is a summary of frequency with which the 250 examiners produced the same item mean deviation score. Observing the bars made it possible to ascertain the distribution of the item mean deviation scores.

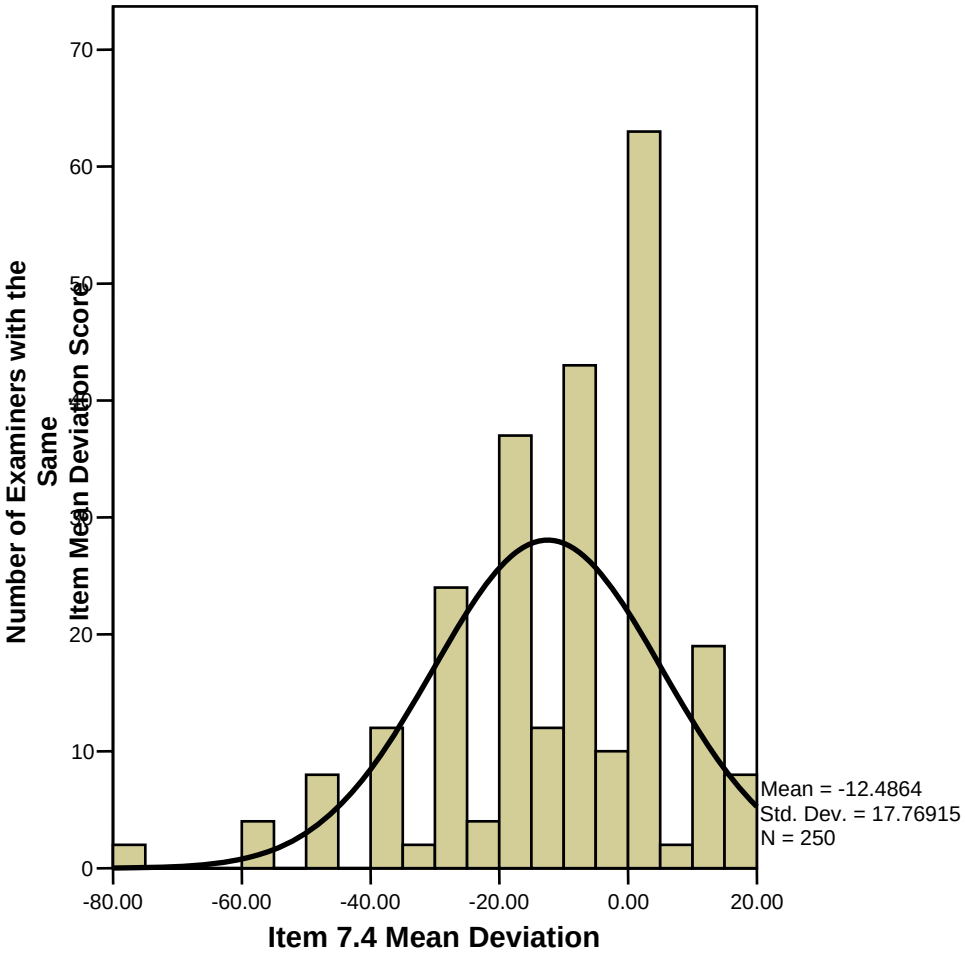

FIGURE 44. Frequency Distribution for Item Mean Deviation Scores for Item 7.4 – Human Resource Results

For Item 7.4, the most prevalent item mean deviation scores were within +/- 20 points of zero. Upon further investigation, it was revealed that approximately 80% of the item mean deviation scores were within +/- 20 points from zero. Therefore, although there was variation overall, the majority of examiners assessed the item closer to the team item consensus score than those who did not. As the scores move farther away from zero, the scores were less stable, as evidenced by the variation in length of the bars on the histogram.

Items 7.1, 7.2, 7.3 and 7.4 were the four items that made up the Results category. Although there was variation within each item, examiners did score relatively consistently across all four. Examiner assessments resulted in greater variation, and lower scores overall, than other categories. Item mean deviation scores ranged from 79% to 82% falling within +/- 20 points of zero for all four items. The similar means and skewness values indicate that examiners assessed human resource focus with a similar approach. Possible causes and implications will be discussed in Chapter V.

Summaries of the means, standard deviations and % of examiners who scored within +/- 20 points from the team item consensus score are presented in Tables 33 – 36. Table 33 is a summary listed in order of the Items. Table 34 is a summary ranked in order of the mean deviation, from smallest to largest. Table 35 is a summary ranked in order of standard deviations from smallest to largest. Table 36 is a summary ranked in order of % of examiners whose item mean deviation scores were within +/- 20 points from zero from highest percentage to lowest percentage.

Observation of Table 33 reveals that Items 7.1 – 7.4, which are the items focused on results, have the largest means. Larger means indicate that examiners had more variation and therefore less consistency in scoring when compared with team item consensus scores. This result may stem from the fact that examiners had to consider overall results by looking at organizations' abilities to deploy continuous improvement strategies throughout the entire organization rather than simply looking at how the organization approached this item area. For example, Item 7.1 – Customer-Focused Results, focused on an organization's ability to deploy the approach each organization outlined for Items 3.1 and 3.2 from the Customer and Market Focus

category (Category 3.0). Larger means in the Results category of 7.0 may indicate that interpreting results is more difficult to assess than other categories (1.0-6.0) where examiners must assess the approach of the organizations.

TABLE 33. Summary of Means, Standard Deviations, and Percent of Examiners Score within +/- 20 Points from Zero for All Items of the TAPE Ranked by Item

Item	Mean	Standard Deviation	% +/- 20 pts from 0
1.1	-6.81	16.67	86%
1.2	-7.57	17.89	85%
2.1	-9.67	17.9	83%
2.2	-9.45	18.05	83%
3.1	-9.52	17.88	84%
3.2	-8.3	16.61	86%
4.1	-6.83	17.04	85%
4.2	-8.69	17.73	83%
5.1	-6.13	15.77	88%
5.2	-5.07	16.07	88%
5.3	-6.35	17.11	88%
6.1	-8.43	17.66	86%
6.2	-10.02	19.1	80%
7.1	-12.12	17.26	80%
7.2	-11.56	17.42	79%
7.3	-11.53	16.15	82%
7.4	-12.49	17.77	80%

Table 34 is a summary ranked in order of mean deviation from smallest to largest. Observation of Table 34 reveals that Category 5.0 – Human Resource Focus, consisting of Items 5.1, 5.2, and 5.3, is a category that examiners seem to have the most consistency in scoring. These three items, all of which fall in the same category, also had the greatest percentage of examiners scoring within +/- 20 points from zero. Items 7.1, 7.2, 7.3, and 7.4, all of which fall in the same category, again appear to be

the most difficult for examiners to score consistently as evidenced by the means with the largest deviation from the team item consensus score. Likewise, the percentage of examiners scoring within +/- 20 points from zero is the smallest compared to the other items.

TABLE 34. Summary of Means, Standard Deviations, and Percent of Examiners Score within +/- 20 Points from Zero for All Items of the TAPE Ranked by Mean

Item	Mean	Standard Deviation	% +/- 20 pts from 0
5.2	-5.07	16.07	88%
5.1	-6.13	15.77	88%
5.3	-6.35	17.11	88%
1.1	-6.81	16.67	86%
4.1	-6.83	17.04	85%
1.2	-7.57	17.89	85%
3.2	-8.3	16.61	86%
6.1	-8.43	17.66	86%
4.2	-8.69	17.73	83%
2.2	-9.45	18.05	83%
3.1	-9.52	17.88	84%
2.1	-9.67	17.9	83%
6.2	-10.02	19.1	80%
7.3	-11.53	16.15	82%
7.2	-11.56	17.42	79%
7.1	-12.12	17.26	80%
7.4	-12.49	17.77	80%

Table 35 is a summary ranked in order of the standard deviation from smallest to largest. Observation of Table 35 reveals that rank ordering the item mean deviation scores by standard deviation does not yield a pattern. It is interesting to note that the standard deviations are relatively stable across all items with a range of less than 4.5

TABLE 35. Summary of Means, Standard Deviations, and Percent of Examiners Score within +/- 20 Points from Zero for All Items of the TAPE Ranked by Standard Deviation

Item	Mean	Standard Deviation	% +/- 20 pts from 0
5.1	-6.13	15.77	88%
5.2	-5.07	16.07	88%
7.3	-11.53	16.15	82%
3.2	-8.3	16.61	86%
1.1	-6.81	16.67	86%
4.1	-6.83	17.04	85%
5.3	-6.35	17.11	88%
7.1	-12.12	17.26	80%
7.2	-11.56	17.42	79%
6.1	-8.43	17.66	86%
4.2	-8.69	17.73	83%
7.4	-12.49	17.77	80%
3.1	-9.52	17.88	84%
1.2	-7.57	17.89	85%
2.1	-9.67	17.9	83%
2.2	-9.45	18.05	83%
6.2	-10.02	19.1	80%

Table 36 is a summary ranked in order of percent of examiners whose item mean deviation scores were within +/- 20 points from zero ranked from highest to lowest. Observation of Table 36 reveals similar patterns as that of Table 34 where items are rank ordered by means. The top three items with the smallest variation from the team item consensus scores are items from Category 5.0 – Human Resource Focus. The items with the greatest variation from the team item consensus scores are items from Category 7.0 – Results and Item 6.2 – Support Processes. This similarity suggests that there is a correlation between means and the percentage of examiners that score within +/- 20 points from zero.

TABLE 36. Summary of Means, Standard Deviations, and Percent of Examiners Score within +/- 20 Points from Zero for All Items of the TAPE Ranked by % within +/- 20 Points from Zero

Item	Mean	Standard Deviation	% +/- 20 pts from 0
5.1	-6.13	15.77	88%
5.2	-5.07	16.07	88%
5.3	-6.35	17.11	88%
3.2	-8.3	16.61	86%
1.1	-6.81	16.67	86%
6.1	-8.43	17.66	86%
4.1	-6.83	17.04	85%
1.2	-7.57	17.89	85%
3.1	-9.52	17.88	84%
4.2	-8.69	17.73	83%
2.1	-9.67	17.9	83%
2.2	-9.45	18.05	83%
7.3	-11.53	16.15	82%
7.1	-12.12	17.26	80%
7.4	-12.49	17.77	80%
6.2	-10.02	19.1	80%
7.2	-11.56	17.42	79%

Summary of Research Question 2

Research Question 2 was a focus on the differences between item mean deviation scores and team item consensus scores. The data disaggregated by each of the 17 items are a revelation that, in general, individual examiners did not produce scores that were consistent with their team's total consensus scores, and that individuals tended to score more critically when working on their own than when coming to consensus as a team, as was indicated by the fact that all item mean deviation scores were below zero. There appeared to be little variation between the standard deviations and no apparent pattern associated with standard deviations across items. Additionally, the ranges falling between minimum scores and maximum scores

revealed little variation and no particular pattern. There was little variation in the skewness values; all values were negative.

With the exception of Item 7.2, which had 79% of examiners assess the item within +/- 20 points from zero, at least 80% of examiners had item mean deviation scores with +/- 20 points from zero for all other items. Given that the average percentage of 85% and the standard deviations are relatively consistent, this may suggest that examiner variation within an item is very acceptable and certainly lies within acceptable bounds. This may further suggest that the examiner assessment processes are basically stable and therefore afford a degree of reliability in the overall assessment.

The pattern that appears to emerge from the analysis for the question is that examiners have the greatest consistency in scoring as compared to the team item consensus score for all items having to do with the Category 5.0 – Human Resource Focus. This consistency could be due to the fact that training around those items is more effective. The consistency could also be attributed to the way that the questions for those items are worded. Organizations could be better at describing their processes and approaches to areas focusing on human resources. Everyone in organizations has some interaction with human resources in their job, which may account for examiners being able to more easily identify with what organizations describe in regard to their human resource strategies.

The other pattern that appears to emerge from the data is that examiners have the greatest variation in scoring as compared to team item consensus score for all items having to do with Category 7.0 – Results. As noted earlier, the results category varies

somewhat from other categories, in that examiners are assessing the results as indicators of organizations' strategies rather than the approach the organization takes. It may be that deployment is more difficult to assess than approach to assess. Deployment is a broader concept which may be more difficult for an organization to describe, therefore making it more difficult for examiners to identify. The more difficult the explanation, the more difficult it would be for a third-party examiner to assess, which would mean that the potential for individuals to vary from each other would be greater.

Research Question 3 – Do item deviation scores vary significantly across the following classifications:

a. *Levels of Examiner Experience*
b. *Sector*
c. *Levels of self-assessment*
d. *Levels of team experience*

Levels of Examiner Experience

Question 3 required a multivariate analysis of variance (MANOVA) because there were multiple items, or dependent variables. The first classification of deviation score variation was observed in relation to levels of examiner experience. Three categories of examiner experience were used as independent variables, including Senior, Returning, and New. Examiners were placed into the Senior category if they had served as examiners for more than 2 consecutive years. Examiners were placed

into the Returning category if they were returning for their second year of service or if they were returning after having not participated for a year or more. Examiners were placed in the New category if they were serving as examiners for the first time. The label of Senior, Returning, or New examiner was assigned by the staff at the Quality Texas Foundation based on their records of examiner participation. The number of examiners in each of the categories is displayed in Table 37.

TABLE 37. Frequency for Levels of Experience

Value Label	Frequency
Senior	54
Return	76
New	11

Descriptive information for deviation scores were obtained for the 17 items embedded in the seven categories of the performance excellence criteria (see Table 38). In all cases, New examiners produced item mean deviation scores that were less consistent with the team item consensus score when compared to Returning or Senior examiners and had greater within variation (note standard deviations for New examiners compared to Senior and Returning examiners).

TABLE 38. Summary of Descriptive Statistics for Levels of Examiner Experience

Item	Experience Year	Mean	Std. Deviation	N
D1.1 – Organizational Leadership	Senior	-.7352	11.39081	54
	Return	-3.2355	14.89355	76
	New	-12.0672	18.12225	119
	Total	-6.9141	16.62153	249
D1.2 – Social Responsibility	Senior	-4.3130	13.07105	54
	Return	-5.1842	18.79483	76
	New	-10.3017	18.69375	119
	Total	-7.4410	17.81103	249
D2.1 – Strategy Development	Senior	-3.3870	12.55342	54
	Return	-8.5921	17.86854	76
	New	-13.2092	19.26907	119
	Total	-9.6699	17.94056	249
D2.2 – Strategy Deployment	Senior	-2.5407	14.63518	54
	Return	-6.1664	16.94676	76
	New	-14.7626	18.74513	119
	Total	-9.4884	18.08468	249
D3.1 – Customer and Market Knowledge	Senior	-5.6037	12.03933	54
	Return	-6.5033	17.87529	76
	New	-13.0618	19.43062	119
	Total	-9.4426	17.87146	249
D3.2 – Customer Relationship and Satisfaction	Senior	-3.4981	11.76604	54
	Return	-5.1493	15.17562	76
	New	-12.5660	18.37997	119
	Total	-8.3357	16.63626	249
D4.1 – Measurement and Analysis of Organizational Performance	Senior	-.1519	12.90420	54
	Return	-6.0717	16.85834	76
	New	-10.4063	17.97246	119
	Total	-6.8594	17.06600	249

TABLE 38. Continued

Item	Experience Year	Mean	Std. Deviation	N
D4.2 – Information and Knowledge Management	Senior	-3.3074	12.78374	54
	Return	-6.6000	17.28453	76
	New	-12.5479	19.18329	119
	Total	-8.7285	17.75269	249
D5.1 – Work Systems	Senior	-2.4111	13.05839	54
	Return	-4.8836	15.10511	76
	New	-8.5903	17.01322	119
	Total	-6.1189	15.79739	249
D5.2 – Employee Learning and Motivation	Senior	.3333	12.62672	54
	Return	-3.6711	16.77012	76
	New	-8.4538	16.37575	119
	Total	-5.0884	16.09874	249
D5.3 – Employee Well-Being and Satisfaction	Senior	-.5611	12.93303	54
	Return	-2.8632	16.08162	76
	New	-11.1723	18.20634	119
	Total	-6.3349	17.14216	249
D6.1 – Value Creation Processes	Senior	-6.1074	15.54063	54
	Return	-6.3974	17.22385	76
	New	-10.8504	18.68492	119
	Total	-8.4627	17.68323	249
D6.2 – Support Processes	Senior	-4.9907	13.03343	54
	Return	-7.6414	17.29833	76
	New	-13.9979	21.57784	119
	Total	-10.1044	19.05677	249
D7.1 – Customer-Focused Results	Senior	-4.4167	11.54462	54
	Return	-11.8257	15.45461	76
	New	-15.9139	19.33882	119
	Total	-12.1727	17.27751	249
D7.2 – Product and Service Results	Senior	-5.7981	13.65669	54
	Return	-9.4276	15.99827	76
	New	-15.5597	18.96990	119
	Total	-11.5711	17.45798	249

TABLE 38. Continued

Item	Experience Year	Mean	Std. Deviation	N
D7.3 – Financial and Market Results	Senior	-3.9648	12.41864	54
	Return	-9.6447	15.00640	76
	New	-16.2613	16.89658	119
	Total	-11.5751	16.16143	249
D7.4 – Human Resource Results	Senior	-5.8222	11.86232	54
	Return	-10.0934	17.75403	76
	New	-17.2277	18.75522	119
	Total	-12.5767	17.74737	249

The multivariate test result was an indicator that there were significant differences (P = .002) across the levels of experience. Based on the results of the multivariate tests (see Table 39), the hypothesis that the vector of means was equal for the levels of experience was rejected.

TABLE 39. Summary of Wilks' Lambda Multivariate Test

Effect	Value	F	Hypothesis df	Error df	P
Wilks'Lambda	.771	1.880	34.000	460.000	.002

As a result, F-tests (between-subjects effects) were run to determine if there were differences for each dependent variable at each level of experience (see Table 40).

The significance for deviation scores were obtained through tests of between-subject effects. A summary of the results is presented in Table 40. All deviation scores were significant at alpha .05 except for Item D6.1 Value Creation Processes

(Process Management category) which had a probability of .125. Strength of association measures as reflected by partial eta squared are shown in Table 40.

Most items showed small results for strength of association measures. However, Items 1.1-Organizational Leadership, 2.2 – Strategy Deployment, 3.2 – Customer Relationship and Satisfaction, 5.3 – Employee Well-Being and Satisfaction, 7.1 – Customer-Focused Results, 7.3 – Financial and Market Results, and 7.4 – Human Resource Results yielded results in the medium range of strength of association measures. This means that knowing an examiner's level of experience is moderately related to the item or dependent variable.

TABLE 40. Summary of Between-Subjects Effects for All Deviation Scores

Dependent Variable	Type III Sum of Squares	df	Mean Square	F	P	Partial Eta Squared (power)
D1.1 – Organizational Leadership	6250.101	2	3125.051	12.346	.000*	.091
D1.2 – Social Responsibility	1889.281	2	944.640	3.026	.050*	.024
D2.1 – Strategy Development	3710.608	2	1855.304	5.997	.003*	.046
D2.2 – Strategy Deployment	6755.518	2	3377.759	11.175	.000*	.083
D3.1 – Customer and Market Knowledge	3011.114	2	1505.557	4.861	.009*	.038
D3.2 – Customer Relationship and Satisfaction	4164.845	2	2082.422	7.946	.000*	.061
D4.1 – Measurement and Analysis of Organizational Performance	3973.761	2	1986.881	7.161	.001*	.055
D4.2 – Information and Knowledge Management	3667.234	2	1833.617	6.055	.003*	.047

TABLE 40. Continued

Dependent Variable	Type III Sum of Squares	df	Mean Square	F	P	Partial Eta Squared (power)
D5.1 – Work Systems	1585.210	2	792.605	3.233	.041*	.026
D5.2 – Employee Learning and Motivation	3087.784	2	1543.892	6.207	.002*	.048
D5.3 – Employee Well-Being and Satisfaction	5500.822	2	2750.411	10.042	.000*	.075
D6.1 – Value Creation Processes	1302.189	2	651.094	2.101	.125	.017
D6.2 – Support Processes	3677.061	2	1838.530	5.236	.006*	.041
D7.1 – Customer-Focused Results	4923.142	2	2461.571	8.762	.000*	.067
D7.2 – Product and Service Results	4041.964	2	2020.982	6.949	.001*	.053
D7.3 – Financial and Market Results	6024.032	2	3012.016	12.612	.000*	.093
D7.4 – Human Resource Results	5506.526	2	2753.263	9.328	.000*	.070

*Significant at the 0.05 level

Since the Fs for differences between levels of experience were significant for all dependent variables except Item 6.1, post hoc tests were run on each of the remaining dependent variables to determine where the differences were across experience levels. In Table 41, a line summary notation of the post hoc test results for each of the 17 deviation scores is utilized.

TABLE 41. Line Notation Summary of Post Hoc Test Results for Overall Item Deviation Scores of the Texas Award for Performance Excellence, 2001-2004

Description	New	Returning	Senior
Item 1.1 Organizational Leadership			
Item 1.2 Social Responsibility			
Item 2.1 Strategy Development			
Item 2.2 Strategy Deployment			
Item 3.1 Customer and Market Knowledge			
Item 3.2 Customer Relationship and Satisfaction			
Item 4.1 Measurement. and Analys. of Org. Perf.			
Item 4.2 Information and Knowledge Mgmt.			
Item 5.1 Work Systems			
Item 5.2 Employee Learning & Motivation			

TABLE 41. Continued

Description	New	Returning	Senior
Item 5.3 Employee Well-Being and Satisfaction			
Item 6.1 Value Creation Processes			
Item 6.2 Support Processes			
Item 7.1 Customer-Focused Results			
Item 7.2 Product and Service Results			
Item 7.3 Financial and Market Results			
Item 7.4 Human Resource Results			

Results for levels of examiner experience were summarized in post hoc tables which placed items with scores falling in the same range into one, two or three homogeneous subsets. Individual post hoc tables can be found in Appendix A. Table 41 was created as a summary representation of the post hoc test results. All results fell in a range that was less than zero so gray bars represent a continuum of values. Therefore, the top row of bars are a representation of values that were farther away from zero in a negative range. Mid-row bars represent items that were approaching

zero, and bottom row bars represent items that had values closest to zero (i.e., more consistent with item consensus scores). In all cases except for Item 7.3, Financial and Market results, there were only two rows of bars because two experience levels fell in the same subset. All post hoc test results yielded examiner levels in the same order; new examiners had scores that were least consistent with the item consensus scores and senior examiners had scores that were most consistent with the item consensus scores. Returning examiners were consistently in between new and senior level examiners.

Out of the 17 items (including item 6.1), 10 had post hoc results indicating new examiners were in one subset farther away from zero (new examiners had significantly greater variation from zero than did returning or senior examiners), while returning and senior examiners were in the same subset, meaning they were basically within the same statistical range regarding consistency with item consensus scores. For example the post hoc results for Item 1.1- Organizational Leadership reflected the following: New examiners fell into one subset with an item mean deviation score of -12.07 while Returning and Senior fell into a second subset with item mean deviation scores of -3.24 and -.74 respectively. On a continuum, such as the one simulated in the line summary notation in Table 41, New examiners' item mean deviation scores are shifted farther to the left to indicate their greater variation from zero. Observing Table 41, it can be ascertained that New examiners fall in one subset that is separated from Returning and Senior examiners who fall into a second subset that is closer to zero.

Item 1.2 – Social Responsibility and item 6.1 – Value Creation Processes had results indicating examiners at all three levels of experience fell into the same subset; meaning that none of the three levels of examiners varied significantly from each other in their item mean deviation scores. Therefore, the bar for each group of examiners was on one level. Items 2.1 – Strategy Development and 5.1 Work Systems had results indicating New and Returning examiners fell into one subset and Returning and Senior examiners fell in a second subset. In these items, Returning examiners had an item mean deviation score that was not significantly different from New examiners or Senior examiners, yet New and Senior examiners' item mean deviation scores did vary significantly. As a result, the two levels of bars overlap in the Returning examiners column. Item 7.3 was the only item where the three levels of experience, New, Returning, and Senior each fell into separate subsets, meaning that all three groups had item mean deviation scores that were significantly different from each other, and Senior examiners had scores closest to the team item consensus score. Given the variation between and among new, returning and senior examiners, it appears that after one year's experience and training as an examiner, the variation gap is mitigated in such a way that returning and senior examiners do not, for most items, vary significantly.

Sectors. The second independent variable was the group of sectors. Overall, there were 34 assessments across the six sectors. The number of examiners in each sector is presented in Table 42.

TABLE 42. Between-Subjects Factors – Number of Examiners across Sectors

Sector	N
Service	39
Manufactory	76
Health Care	28
Small Organization	40
Education	45
Public	22

Because results of the multivariate tests for sector showed that sectors were significantly different from each other overall (see Table 43), it was necessary to probe items by each sector.

TABLE 43. Summary of Multivariate Tests for Sector

	Value	F	Hypothesis df	Error df	P	Partial Eta Squared
Wilks' Lambda	.423	2.538	85.000	1106.609	.000	.158

Table 44 is a summary of the deviation scores that indicated significant differences (P <.01 or P < .05) and shows that the dependent variables D3.1 - Customer and Market Focus (P = .049), D5.1 – Work Systems (sig. .043), D5.2 – Employee Learning and Motivation (P = .033), D5.3 – Employee Well-Being and Satisfaction (P = .001), D6.1 – Value Creation Processes (P = .023) and D6.2 –

Support Processes (P = .037) were the items in which item mean deviation scores were significantly different from each other. This is interesting that Items 5.1, 5.2 and 5.3 varied significantly between and among sectors when viewed in light of Items 5.1, 5.2, and 5.3 having the least mean deviation scores overall (see Table 34).

TABLE 44. Summary of Tests of Between-Subjects Effects for Sector

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	P	Partial Eta Squared
Sector	D1.1 – Organizational Leadership	2666.661	5	533.332	1.955	.086	.039
	D1.2 – Social Responsibility	406.920	5	81.384	.250	.939	.005
	D2.1 – Strategy Development	2514.520	5	502.904	1.587	.164	.032
	D2.2 – Strategy Deployment	441.135	5	88.227	.267	.931	.005
	D3.1 – Customer and Market Knowledge	3521.960	5	704.392	2.258	.049*	.044
	D3.2 – Customer Relationship and Satisfaction	2916.644	5	583.329	2.163	.059	.042
	D4.1 – Measurement and Analysis of Organizational Performance	2059.714	5	411.943	1.431	.213	.028
	D4.2 – Information and Knowledge Management	3279.921	5	655.984	2.135	.062	.042
	D5.1 – Work Systems	2825.144	5	565.029	2.334	.043*	.046
	D5.2 – Employee Learning and Motivation	3092.914	5	618.583	2.466	.033*	.048
	D5.3 – Employee Well-Being and Satisfaction	5916.790	5	1183.358	4.311	.001*	.081
	D6.1 – Value Creation Processes	4001.883	5	800.377	2.653	.023*	.052
	D6.2 – Support Processes	4260.331	5	852.066	2.412	.037*	.047
	D7.1 – Customer-Focused Results	855.332	5	171.066	.569	.724	.012
	D7.2 – Product and Service Results	718.632	5	143.726	.468	.800	.010
	D7.3 – Financial and Market Results	1365.860	5	273.172	1.049	.390	.021
	D7.4 – Human Resource Results	3047.758	5	609.552	1.968	.084	.039

*Significant at 0.05 level

Item 3.1 in the Customer and Market Focus category was the item with the greatest significance. The Post Hoc test results for this item are summarized in Table 45. Of the six sectors, the Public sector and Service sector varied the most with -15.2273 and -2.0513 respectively, which indicates the examiners did not have consistency when rating these sectors.

TABLE 45. Post Hoc Test Results by Sector for Item 3.1 Customer and Market Knowledge for the Texas Award for Performance Excellence

Sector	N	Subset	
		1	2
Public	22	-15.2273	
Manufacturing	76	-11.7132	-11.7132
Education	45	-11.4000	-11.4000
Health Care	28	-8.3929	-8.3929
Small Organization	40	-8.2000	-8.2000
Service	39		-2.0513
Sig.		.556	.070

Table 46 is a line notation summary table of the six items that were significant. Individual post hoc tables are located in Appendix B. As in Table 41, the gray bars represent a continuum of values up to zero. The continuum moving from left to right indicates values farthest to the right are closest to zero (i.e. more consistent with item consensus scores). In each of the six post hoc test results for items that were significant, sectors fell in a range of two homogeneous subsets. Therefore, there are

only two rows of bars. The top rows are a representation of values that were in one subset based on the fact that the item mean deviation scores in the identified items were not significantly different from each other, and bottom row of bars represent the second set of items that had item mean deviation scores were not significantly different from each other. Observing items at the farthest ends of the two rows (i.e. the items that do not overlap with each other) makes it possible to determine those items that were significantly different from each other. For example, the set of rows for Item 3.1 represent results indicating Public and Service sectors are significantly different from each other, yet the other sectors are not (indicated by the overlapping bars). Since post hoc test results yielded sectors in different orders; it was necessary to place sector identifiers on the bars to show where there was overlap.

TABLE 46. Summary of Post Hoc Test Results for Sectors of the Texas Award for Performance Excellence, 2001-2004

Sector	Distribution					
Item 3.1 Customer and Market Knowledge	P	M	E	HC	SO	
		M	E	HC	SO	S
Item 5.1 Work Systems	P	SO	M	HC	E	
			M	HC	E	S
Item 5.2 Employee Learning and Motivations	P	SO	E	M	S	
		SO	E	M	S	HC
Item 5.3 Employee Well-being and Satisfaction	P	E	M	SO		
			M	SO	S	
				SO	S	HC
Item 6.1 Value Creation Processes	E	P	SO	S	HC	
		P	SO	S	HC	M

TABLE 46. Continued

Sector	Distribution					
Item 6.2 Support Processes and Operational Planning	E	P	M	S	SO	
		P	M	S	SO	HC

P = Public, M = Manufacturer, E = Education, HC = Health Care, SO = Small Organization, S = Service

Sectors with item mean deviation scores with the greatest variation from the team item consensus score were Public and Education sectors. This is not necessarily surprising given these sectors' relative immaturity in participating in the TAPE process. Sectors that were closest to the team item mean consensus score were service, health care, and manufacturing organizations. Given that the Public sector had the highest deviation from the mean in all but Items 6.1 and 6.2, it may be that the Public sector has the greatest difficulty in understanding and applying the Generic TAPE Criteria since it is based on a primarily business model. Education and Health Care have their own tailored criteria.

It is interesting to note that the Education sector appears to have the most difficulty and variation with assessing Items 6.1 – Value Creation Processes and Item 6.2 – Support Processes. The accountability system in K-12 education may focus educators so intently on results or that processes are so ill-defined or vary so greatly that educators have difficulty in clearly identifying and articulating value creation and support processes. On the other hand, the Service and Health Care sectors are both relatively new to the TAPE process and yet they show the least variation form the mean for these items

Levels of self-assessment. The third independent variable was levels of self-assessment. There were five levels of self-assessment from which examiners could choose in order to rate their own self-confidence in relation to the actual organization to which they were assigned. Examiners were asked to rate themselves using a Likert-type scale of 1 to 5 on their confidence level regarding their ability to accurately assess their assigned organization. The researcher assigned the numbers into the following labels:

- 5: Highly confident
- 4: Confident
- 3: Somewhat Confident
- 2: Slightly confident
- 1: Not confident

Table 47 is a summary of how confident examiners were in their own ability to accurately assess an organization's performance based on the Criteria for Performance Excellence.

TABLE 47. Frequency for Levels of Self-Assessment

Confidence Level	N
Not confident	0
Slightly confident	16
Somewhat confident	60
Confident	95
Highly confident	71

There were no instances where an examiner rated himself or herself as "not confident." Therefore, the label "not confident" was dropped from the final analysis. Eight examiners did not enter a self-assessment code. Consequently, the total data set for this category is N=242. A summary of the overall outcome of mean ratings of items by levels of self-assessment is presented in Table 48.

TABLE 48. Summary of Examiners' Self-Assessment Rating for Each Item of the TAPE

	Self Assessment	Mean	Std. Deviation	N
D1.1 – Organizational Leadership	Slightly confident	-2.2875	16.44027	16
	Somewhat confident	-6.0683	19.53530	60
	Confident	-7.2211	15.64848	95
	Highly confident	-7.3690	16.01687	71
	Total	-6.6525	16.78999	242
D1.2 – Social Responsibility	Slightly confident	-2.4375	14.34326	16
	Somewhat confident	-6.9167	18.06607	60
	Confident	-8.4874	17.66127	95
	Highly confident	-6.5662	18.66065	71
	Total	-7.1343	17.82710	242
D2.1 – Strategy Development	Slightly confident	-8.7500	15.43805	16
	Somewhat confident	-8.1167	17.91703	60
	Confident	-11.4032	18.21674	95
	Highly confident	-9.5521	17.49625	71
	Total	-9.8698	17.70814	242
D2.2 – Strategy Deployment	Slightly confident	-5.2063	15.20353	16
	Somewhat confident	-8.0242	19.62612	60
	Confident	-13.6953	17.77224	95
	Highly confident	-7.1423	16.45594	71
	Total	-9.8054	17.91044	242
D3.1 – Customer and Market Knowledge	Slightly confident	-5.4375	14.55092	16
	Somewhat confident	-10.3125	20.67698	60
	Confident	-9.7258	17.07693	95
	Highly confident	-9.2789	17.64122	71
	Total	-9.4566	17.97988	242

TABLE 48. Continued

	Self Assessment	Mean	Std. Deviation	N
D3.2 – Customer Relationship and Satisfaction	Slightly confident	-1.9188	12.28165	16
	Somewhat confident	-6.9925	18.82359	60
	Confident	-9.3542	15.56816	95
	Highly confident	-9.1507	16.86904	71
	Total	-8.2174	16.64611	242
D4.1 – Measurement and Analysis of Organizational Performance	Slightly confident	-3.0250	13.20745	16
	Somewhat confident	-8.6225	16.83673	60
	Confident	-6.6353	17.82374	95
	Highly confident	-6.5577	17.28810	71
	Total	-6.8665	17.11064	242
D4.2 – Information and Knowledge Management	Slightly confident	-8.7063	14.33948	16
	Somewhat confident	-7.9600	19.02429	60
	Confident	-9.1568	17.49885	95
	Highly confident	-9.2338	18.64461	71
	Total	-8.8529	17.94670	242
D5.1 – Work Systems	Slightly confident	-1.4562	15.44228	16
	Somewhat confident	-6.8158	15.85864	60
	Confident	-6.6268	15.37819	95
	Highly confident	-6.9028	16.71425	71
	Total	-6.4128	15.86052	242
D5.2 – Employee Learning and Motivation	Slightly confident	.4375	13.82254	16
	Somewhat confident	-7.4000	16.79205	60
	Confident	-4.5895	16.89674	95
	Highly confident	-5.7887	15.28951	71
	Total	-5.3058	16.23982	242
D5.3 – Employee Well-Being and Satisfaction	Slightly confident	-1.4188	17.15942	16
	Somewhat confident	-7.7550	17.40320	60
	Confident	-5.6653	17.67899	95
	Highly confident	-7.7592	16.74427	71
	Total	-6.5169	17.28186	242
D6.1 – Value Creation Processes	Slightly confident	-2.9312	17.37994	16
	Somewhat confident	-6.3683	17.11009	60
	Confident	-10.1926	18.62424	95
	Highly confident	-9.3789	17.38268	71
	Total	-8.5256	17.83456	242

TABLE 48. Continued

	Self Assessment	Mean	Std. Deviation	N
D6.2 – Support Processes	Slightly confident	-5.0625	18.28467	16
	Somewhat confident	-9.2875	17.33841	60
	Confident	-11.7237	19.48293	95
	Highly confident	-9.1408	20.89219	71
	Total	-9.9215	19.29677	242
D7.1 – Customer-Focused Results	Slightly confident	-10.4375	18.19512	16
	Somewhat confident	-11.8542	17.72831	60
	Confident	-14.1132	18.21371	95
	Highly confident	-9.9437	15.11906	71
	Total	-12.0868	17.21920	242
D7.2 – Product and Service Results	Slightly confident	-4.3750	16.62077	16
	Somewhat confident	-9.7917	15.35106	60
	Confident	-13.1916	19.17749	95
	Highly confident	-11.3915	16.76096	71
	Total	-11.2376	17.46684	242
D7.3 – Financial and Market Results	Slightly confident	-5.1875	15.44547	16
	Somewhat confident	-9.6667	14.65151	60
	Confident	-13.6074	17.27212	95
	Highly confident	-11.8704	16.22395	71
	Total	-11.5640	16.29366	242
D7.4 – Human Resource Results	Slightly confident	-6.3125	16.82743	16
	Somewhat confident	-12.4283	18.16491	60
	Confident	-13.6642	18.35888	95
	Highly confident	-12.3493	17.37358	71
	Total	-12.4860	17.90593	242

Table 49 is a summary of total means ranked from largest to smallest means. It is interesting to note the a similar pattern as was seen in Table 34 where items from Category 5.0 had means that were closest to zero and items from Category 7.0 has means what were farthest from zero.

TABLE 49. Summary of Examiners' Self-Assessment Rating for Each Item of the TAPE Ranked from Largest to Smallest

Item	Self Assessment	Mean	Std. Deviation	N
D5.2 – Employee Learning and Motivation	Total	-5.3058	16.23982	242
D5.1 – Work Systems	Total	-6.4128	15.86052	242
D5.3 – Employee Well-Being and Satisfaction	Total	-6.5169	17.28186	242
D1.1 – Organizational Leadership	Total	-6.6525	16.78999	242
D4.1 – Measurement and Analysis of Organizational Performance	Total	-6.8665	17.11064	242
D1.2 – Social Responsibility	Total	-7.1343	17.82710	242
D3.2 – Customer Relationship and Satisfaction	Total	-8.2174	16.64611	242
D6.1 – Value Creation Processes	Total	-8.5256	17.83456	242
D4.2 – Information and Knowledge Management	Total	-8.8529	17.94670	242
D3.1 – Customer and Market Knowledge	Total	-9.4566	17.97988	242
D2.2 – Strategy Deployment	Total	-9.8054	17.91044	242
D2.1 – Strategy Development	Total	-9.8698	17.70814	242
D6.2 – Support Processes	Total	-9.9215	19.29677	242
D7.2 – Product and Service Results	Total	-11.2376	17.46684	242
D7.3 – Financial and Market Results	Total	-11.5640	16.29366	242
D7.1 – Customer-Focused Results	Total	-12.0868	17.21920	242
D7.4 – Human Resource Results	Total	-12.4860	17.90593	242

There was no significant difference between examiners with different self-assessment levels, as evidenced by the results found in Table 50. Therefore, no post hocs were run and no further analysis was warranted.

TABLE 50. Summary of Wilks' Lambda Multivariate Tests for Self-assessment Levels of Examiners

Effect	Value	F	Hypoth. df	Error df	Sig.	Partial Eta Sqd.
Wilks' Lambda	.828	.850	51.000	661.735	.761	.061

Levels of team experience. The fourth and final independent variable was levels of team experience. Teams were placed into 1 of 3 categories by the researcher based on the make up of examiner experience levels on the team. The three levels of team experience were:

- New
- Average
- Senior

New teams consisted of more than 50% New examiners. Senior teams consisted of more than 50% Returning and Senior examiners. Average teams were split evenly with 50% New examiners and 50% Returning and Senior examiners. Returning and Senior examiners were grouped together based on the fact that both groups had prior experience serving as examiners and would have similar influence on the outcome of organizational assessments. Table 51 is a summary of frequencies of examiners across team experience levels. There were a greater number of teams classified as senior level.

TABLE 51. Frequency for Number of Examiners by Team Experience Level

Description		Value Label	N
Team Experience Levels	1	New Team	74
	2	Average Team	50
	3	Senior Team	126

Table 52 is a summary of the descriptive statistics for means and standard deviation for levels of team experience. Observing Table 52, one can ascertain that, to a certain degree, senior-level teams scored most consistently with the item mean deviation score. However, average-level teams scored more consistently with the item mean deviation score for Items 1.1 – Organizational Leadership, 2.2-Strategy Deployment, and 4.1- Measurement and Analysis of Organizational Performance. New teams scored more consistently than other teams on Items 4.2 – Information and Knowledge Management and Item 5.2 – Employee Learning and Motivation.

TABLE 52. Summary of Descriptive Statistics Means and Standard Deviations of Levels of Team Experience

ITEM	**New** N=74		**Average** N=50		**Senior** N=126		**Total** 250	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1.1 Organizational Leadership	-8.97	15.88	-4.90	16.40	-6.29	17.22	-6.81	16.68
1.2 Social Responsibility	-8.64	15.32	-8.44	18.76	-6.60	19.00	-7.57	17.89
2.1 Strategy Development	-11.66	17.15	-10.24	16.75	-8.28	18.77	-9.67	17.90
2.2 Strategy Deployment	-11.42	16.27	-8.30	21.25	-8.75	17.74	-9.45	18.10

TABLE 52. Continued

ITEM	**New** N=74		**Average** N=50		**Senior** N=126		**Total** 250	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3.1 Customer Market Knowledge	-10.38	18.02	-13.50	17.25	-7.44	17.88	-9.52	17.88
3.2 Customer Relationship and Satisfaction	-8.84	16.56	-10.16	14.93	-7.25	17.30	-8.30	16.61
4.1 Measurement and Analysis of Org. Performance	-7.03	15.11	-5.60	17.45	-7.21	18.02	-6.83	17.04
4.2 Information and Knowledge Management	-8.15	15.07	-8.55	19.11	-9.10	18.70	-8.70	17.73
5.1 Work Systems	-6.10	15.47	-6.60	17.65	-5.97	15.27	-6.13	15.77
5.2 Employee Learning and Motivation	-4.51	15.10	-7.80	18.44	-4.31	15.64	-5.10	16.10
5.3 Employee Well-Being and Satisfaction	-8.32	17.42	-12.40	18. 01	-2.80	15.80	-6.35	17.11
6.1 Value Creation Processes	-9.10	18.49	-10.30	17.45	-7.32	17.29	-8.43	17.70
6.2 Support Processes	-9.54	19.32	-11.30	21.60	-9.80	17.95	-10.02	19.10
7.1 Customer-Focused Results	-13.24	19.68	-9.40	16.21	-12.55	16.14	-12.12	17.26
7.2 Product and Service Results	-14.10	18.80	-9.80	15.92	-10.80	17.14	-11.56	17.42
7.3 Financial and Market Results	-12.62	16.02	-12.60	17.36	-10.46	15.78	-11.53	16.15
7.4 Human Resource Results	-14.28	15.57	-14.80	17.76	-10.51	18.90	-12.49	17.77

Lowest means were shaded for each item along with the total item mean deviation for each item in an effort to identify a pattern in Table 52. As indicated by the shading, there does not appear to be a consistent or predictable pattern with regard to lower means and levels of team experience, except to note that, as previously stated,

senior-level teams tend to have less variation from zero for most items. Results in this table support the earlier finding that new examiners tend to have higher means and standard deviations than do returning or senior examiners.

Table 53 is a summary of the Wilks' Lambda multivariate test result which resulted in an F ratio indicating that there were significant differences between team experience levels overall. The effect size was in the medium range at .108. Post hoc tests were run for each dependent variable revealing that only Item 5.3 out of the other 17 items was significant.

TABLE 53. Summary of Wilks' Lambda Multivariate Test for Team Experience Levels

Effect	Value	F	Hypothesis df	Error df	P	Partial Eta Squared
Wilks' Lambda	.796	1.641[b]	34.000	462.000	.014	.108

Table 54 is a summary of the post hoc test results for Item 5.3. It appears that average-level teams and new teams are alike, in that they exhibit greater deviation from zero. Senior-level teams are in a class by themselves as they exhibit a much smaller deviation (-2.79) and are and, therefore, closer to zero.

TABLE 54. Summary of Post Hoc Test Results for Item 5.3 – Employee Well-Being and Satisfaction

Team Experience Levels	N	Subset	
		1	2
Average team	50	-12.3960	
New team	74	-8.3243	
Senior team	126		-2.7905
Sig.		.185	1.000

Table 55 is a summary of the items that were significant by independent variable. It appears that item 5.3 might benefit from additional training given there is significant variation in 3 of the 4 dependent variables.

TABLE 55. Summary of Independent Variables Showing Significance by Item

Item	Levels of Experience	Sector	Levels of Self-Assessment	Levels of Team Experience
1.1	X			
1.2				
2.1				
2.2	X			
3.1		X		
3.2	X			
4.1				
4.2				
5.1		X		
5.2		X		
5.3	X	X		X
6.1		X		
6.2		X		
7.1	X			
7.2				
7.3	X			
7.4	X			

Summary of research question 3. Research Question 3 focused on the differences in item deviation scores by levels of examiner experience, sector, levels of self-assessment and levels of team experience. Because there were multiple dependent variables, a multivariate analysis of variance (MANOVA) was utilized along with univariate F tests (ANOVA)and post hoc tests.

The multivariate test result was an indicator that there were significant differences (P = .002) across the levels of experience. F-tests revealed that Items 1.1-Organizational Leadership, 2.2 – Strategy Deployment, 3.2 – Customer Relationship and Satisfaction, 5.3 – Employee Well-Being and Satisfaction, 7.1 – Customer-Focused Results, 7.3 – Financial and Market Results, and 7.4 – Human Resource Results yielded results in the medium range of strength of association measures. This meant that knowing an examiner's level of experience was moderately related to the item or dependent variable. Next, post hoc tests were run and results were summarized in the line summary notation in Table 41. Experience levels resulted in the same order for all items; New examiners had the most variation from the team item consensus score and Senior examiners exhibited the least variation from the team item consensus score. Item 7.3 was the only unique item that had post hoc results which revealed that all examiners were significantly different from each other. So, while it was determined that level of examiner experience did have a mild effect on examiners' consistency with team item consensus score, it was not possible to determine a repeating or predictable pattern.

It was determined, based on multivariate test results, that item deviation scores were significantly different from each other when looking across sectors. Univariate

F-test results made it possible to determine that 3.1 - Customer and Market Focus (P = .049), 5.1 – Work Systems (P = .043), 5.2 – Employee Learning and Motivation (P = .033), 5.3 – Employee Well-Being and Satisfaction (P = .001), 6.1 – Value Creation Processes (P = .023) and 6.2 – Support Processes (P = .037) were the items in which item deviation scores were significantly different from each other. Public and Education organizations appeared to have the greatest amount of variation from other sectors. Service and Health Care organizations appeared to have the least amount of variation compared to other sectors. Small Organizations and Manufacturing organizations appeared to fall somewhere in the middle, based on post hoc test results.

Levels of self-assessment turned out not to have any significant effect on examiners' scoring consistency based on multivariate test results. This is good news for the TAPE organization. Since individual staff members in the TAPE office assign examiners to teams and organizations, this result may indicate that their process of placing examiners in teams based in experience in a particular sector is working.

Levels of team experience were also analyzed to see if there were significant differences in consistency of scoring. Descriptive statistics in Table 52 were arranged and shaded in such a way as to identify any patterns or consistencies of means based on level of team experience. Although no pattern emerged, it was possible to ascertain that, in general, teams with a senior-level of experience tended to have smaller means and therefore, more consistency in scoring when compared to the team item consensus scores. Based on the F-test and post hoc test results, item 5.3 – Employee Well-Being and Satisfaction was the only item where there was a

significant difference in scoring consistency of teams. Senior-level teams had a much lower mean (-2.79) than New or Average teams.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This study was designed to add to existing research on third-party assessment by looking at the ability of third-party examiners to assess whether or not organizations successfully implemented continuous improvement strategies based on the Criteria of the Texas Award for Performance Excellence. The Texas Award for Performance Excellence (TAPE) is given each year by the Quality Texas Foundation and recognizes organizations that demonstrate superior performance as it is defined by customer satisfaction and continuous improvement (About Quality Texas, n.d.). The TAPE is a state level award for quality that uses the same criteria as the Malcolm Baldrige National Quality Award for Performance Excellence. The researcher proposed to determine the scoring stability of examiners' scores for the Texas Award for Performance Excellence through the analysis of raw score data collected by the Quality Texas Foundation between the years of 2001 to 2004.

Three research questions addressed this purpose:

1. Is the mean of the deviations of individual total scores from team total consensus scores equal to zero?
2. Is the mean of the deviations of individual item scores from team item consensus scores equal to zero?
3. Do item deviation scores vary across the following classifications?
    a. Levels of Examiner Experience
    b. Sector
    c. Levels of Self-Assessment

d. Levels of Team Experience

The population for this study included the total 250 examiners on 34 teams who examined all applicant organizations for the Texas Award for Performance Excellence from 2001 to 2004. For the purpose of this study, the researcher maintained the following assumptions:

The statistical analyses accurately reflected the consistency in examiners' scoring and the effects of levels of examiner experience, levels of team experience, sector, and levels of self-assessment on examiners' scores.

The interpretation of the data collected would accurately reflect what it was intended to reflect.

This study is the only known study using data from the Texas Award for Performance Excellence scoring outcomes. Because it utilizes the total population of data from the four years included in this study, descriptive statistics were the primary method of analysis. In addition, multivariate analysis of variance was used to analyze the data for Research Question 3.

As discussed in Chapter I of this document, the limitations of this study included the following:

1. The scope of this study is limited to the four years of data collected on the Texas Award for Performance Excellence.
2. The consensus score is used as the "true score" against which other scores are compared.
3. The makeup of examiners and the ratio of experience levels are not consistent across the four years.

4. The training material and activities varied each year according to the changes in the TAPE criteria and the individual staff members who delivered training.
5. Small changes were occasionally made to the award criteria and therefore are slightly different in some categories from year to year.
6. Each team rates a different organization each year.
7. Organizations applying for the TAPE are at different experience levels of quality management and organizational self-assessment.
8. Findings from this study may not be generalized to any other quality award.

The review of the literature supported the premise that it cannot be concluded that organizations are judged consistently over time and across sectors and categories of the quality awards. This study supports the premise that further research needs to be conducted on the effects for training on performance of quality award examiners and variation in examiner scoring. The literature review revealed that, since 1991, state and local quality award programs, most modeled after the Malcolm Baldrige National Quality Award program, have grown from fewer than 10 programs to more than 80 in at least 41 states and that since 1988, more than 1000 applications have been submitted for the Baldrige Award from a variety of types and sizes of organizations.

The literature also revealed that the popularity and influence of quality awards is built on a foundational belief that third-party examiners responsible for assessing organizations can consistently and accurately determine, based on the organization's self assessment, which organizations meet or exceed the established criteria setting it

apart from other applicants. Winning organizations receive recognition as leaders in achieving performance excellence. Once they receive the tremendous accolades and publicity that come with winning the award, they are hence forth regarded as exemplars of how to implement quality principles. One troubling fact exists, however—empirical evidence validating the ability of third-party examiners' to accurately assess organizations is remarkably scarce

Findings from this study inform the Quality Texas Foundation about the stability of examiner scoring for the Texas Award for Performance Excellence Program. Additionally, the results of this study provide some insight into what influences examiners' scores which can lead to improved examiner training.

Improved training could result in increased accuracy and objectivity where examiners are able to consistently identify strengths and opportunities for improvement within organizations, thereby increasing the reliability of the assessment process. When a level of stability can be established for examiners' scores on assessments, there can be more certainty that differences in organizational assessments are not a function of examiner differences, and that organizations applying for the TAPE are evaluated in a consistent manner.

Those responsible for training within organizations, specifically Human Resource Development (HRD) specialists, must continually seek to improve their planning and development of training programs. Results of this study have the potential to inform HRD specialists about what impacts the way third-party examiners of the Texas Award for Performance Excellence view organizations. Consequently, HRD specialists may be able to help leaders and managers of Texas organizations better

understand how to produce clear and effective organizational self-assessment documents to gain accurate and reliable examinations.

Summary of Findings

The key findings of this study suggest that examiners of the Texas Award for Performance Excellence do not score consistently with the team consensus score when working independently. In general, examiners assess more leniently when looking at the organization as a whole, and more critically when looking at each item. The results of this study yielded the following key findings related to each of the three research questions.

Research Question 1

Research Question 1 was, “Is the mean of the deviations of individual total scores from total consensus score equal to zero?” The answer to the research question was, no. Examiners tended to score leniently when working independently than when working to come to consensus and therefore, the team mean deviation scores were consistently higher than the total consensus scores for the teams. There were no apparent patterns of consistency in scoring within the four years of the study. Observing team mean deviation scores from one year to the next, however revealed a possible trend. As each year passed, the team mean deviation scores grew smaller as did the standard deviations.

If the results are, in fact, a trend, they may suggest that examiner training is improving. They may also suggest that as understanding of the TAPE and the

elements of continuous improvement are better understood, organizations themselves may be doing a better job of implementing continuous improvement. Organizations may be doing a better job of completing TAPE applications making it possible for examiners to assess more consistently. More effective training of examiners could be what is leading to more consistency in scoring and resulting in scores being more consistent with the total consensus score. It is this possibility that was the focus of this study.

Research Question 2

Research Question 2 asked, “Is the mean of the deviations of individual item scores from team item consensus scores equal to zero? The answer to Research Question 2 was, no. Examiners tended to score more stringently or critically when assessing items independently than when coming to consensus as a team. All item mean deviation scores were below zero. There appeared to be little variation between the standard deviations and no apparent pattern associated with standard deviations across items. Additionally, the ranges falling between minimum scores and maximum scores revealed little variation and no particular pattern. There was little variation in the skewness values; all values were negative. Roughly 80% of all examiners had mean deviation scores that were within +/- 20 points from zero so, although there appeared to be little variation in consistency of examiners scores as compared to consensus scores, it did appear that examiners were scoring consistently with each other.

The two categories with items containing the largest and smallest variation were category 7.0 – Results and category 5.0 – Human Resource Focus. The Results category was somewhat different from other categories, in that examiners assessed the results due to the approach and deployment of the organizations' strategies rather than the approach the organization took. Results is a broader concept which may be more difficult for an organization to describe, therefore making it more difficult for examiners to identify. The more difficult the explanation, the more difficult it would be for a third-party examiner to assess, which would mean that the potential for individuals to vary from each other would be greater. The Human Resource Focus category had the smallest amount of variation as compared to the team item consensus scores. There are several possibilities that could explain why examiners seemed to have less variation where this item was concerned. Training around those items may have been more effective or questions for those items may have been better written. Organizations may have been better at describing their processes and approaches to areas focusing on human resources. Examiners may have been better able to identify with what organizations attempted to describe in regard to their human resource strategies. There were not enough data to be able to discern a definite reason for the large and small variation. This will be recommended for further study.

Research Question 3

Research Question 3 was, "Do item deviation scores vary significantly across the following classifications:

a. Levels of Examiner Experience

b. Sectors

c. Levels of self-assessment

d. Levels of team experience”

Levels of examiner experience. Results from a multivariate analysis of variance indicated that there were significant differences in levels of examiner experience in regard to whether or not there was variation in item deviation scores. Results of F-tests revealed that there was a medium range of strength of association measures for Items 1.1, 2.2, 3.2, 5.3, 7.1, and 7.3 which meant that knowing an examiner’s level of experience was moderately related to an item deviation score. Post hoc tests were analyzed and made it possible to determine that New examiners had the most variation and Senior examiners had the least variation in terms of item deviation scores when compared to team item consensus scores.

Sectors. Results from a multivariate analysis of variance indicated that item deviation scores were significantly different from each other when looking across sectors. Results of F-tests and post hoc tests made it possible to determine that examiners in the Public and Education sector appeared to have the greatest amount of variation from examiners in other sectors. Examiners in the Service and Health Care organizations appeared to have the least amount of variation compared to examiners in other sectors. Small Organizations and Manufacturing organizations appeared to fall somewhere in the middle, based on post hoc test results.

Levels of self-assessment. Tests on levels of self-assessment yielded no indications that there was a significant effect on examiners' scoring consistency. Since individual staff members in the TAPE office assign examiners to teams and organizations, this result may indicate that their process of placing examiners in teams based in experience in a particular sector is working and needs little or less scrutiny when determining areas for improvement in the examiner training process.

Levels of team experience. In general, senior-level teams tended to have smaller means when compared to the team item consensus scores and therefore, more consistency in scoring. Based on the F-test and post hoc test results, item 5.3 – Employee Well-Being and Satisfaction was the only item where there was a significant difference in scoring consistency of teams. Post hoc test results revealed a significant difference between New and Average teams (greater variation) compared to Senior teams (less variation). More data would need to be collected and analyzed to determine the causes of this outcome.

Conclusions

Based on the framework of the three research questions, the limitations of the data and a review of the literature, some conclusions can be drawn concerning the consistency in examiners' scoring for the Texas Award for Performance Excellence.

- Finding stability of examiners' scores is a problem due to the fact that organizations and teams do not repeat.

- Evidence in this study is insufficient to make a determination to what degree the various factors influences examiners' scoring consistency. Therefore, more data needs to be collected in such a way as to make it possible to identify variables that impact consistency of examiner scores.
- The Quality Texas Foundation needs to follow its own philosophy of continuous improvement through measurable data by designing training that would allow for collection of longitudinal data that repeats across examiners.
- Approximately 85% of examiners scored within +/- 20 points from zero. Given the fact that the scoring bands for the TAPE (see Table 4) are in increments of 10, examiners were never more than one band away from scoring consistently with the team consensus score. The Quality Texas Foundation needs to take steps to determine an acceptable level of variation in examiner scoring to further mitigate excessive variation.

Recommendations for Practice

Several recommendations for practice were developed over the course of this study and through the conclusions derived. These recommendations may serve to enhance examiner training and effectiveness of examiners for the Texas Award for Performance Excellence.

1. Utilize the Plan, Do, Study, Act (PDSA) tool by creating a repeatable process that is controllable, measurable, and standards-based. For example, examiners currently complete a case study assessment prior to attending training. Build a tracking tool that would allow Quality Texas

staff to maintain a database that keeps track of examiner's score compared to the ideal (which would be possible to generate in a case study). Then track their variation in individual scores from their team consensus scores during the actual assessment. Finally, have them work through a follow-up case study and track scores against an ideal score for a third measure. Then provide feedback to the examiner on strengths and opportunities for improvement and retrain based on outcomes. This would give the Quality Texas office an idea of the progress each examiner makes within a year and would help to isolate problem areas and items which could inform training and make it more strategic.

2. Consider building and maintaining a database that would allow longitudinal tracking of individual examiners' scores to determine possible causes or trends in variation.
3. Create pre-post tests that would allow Quality Texas to collect a shorter/quicker measure of examiner competence and accuracy in scoring and would take less time than the recommendation mentioned above.
4. Develop and administer surveys for examiners that focus on what influenced them during the consensus meeting. Compare their responses to their individual deviations from consensus scores during assessments. Look for trends within teams to determine how scores fall and to observe the change in directionality and numerical value in scores from individual to consensus.

5. Determine an acceptable level of variation from consensus and track examiner scores over time along with training strategies to determine the impact of training on examiner accuracy.
6. Determine an acceptable percentage of examiners who score within an acceptable range of the consensus score to track progress and inform training.
7. Assign multiple teams to one organization to track consistency in examiner scoring.

Recommendation for Future Research

While the data for this study included the total population of examiner scores for the four years over which the study spanned, it was not enough data to reach strong statistical conclusions about what influenced consistency in examiner scoring. In order to verify and further extend understanding of what influences examiners' scores and what impact training has on examiners, the following recommendations for next steps for research and analysis may be useful.

1. Replicate the study using a larger data set.
2. Consider using the Malcolm Baldrige National Quality Award data and compare it to Texas Award for Performance Excellence data to create a larger data set and generate comparisons.
3. Survey examiners or conduct a qualitative analysis to ascertain perceptions of what influences them during consensus meetings then compare to

their actual individual data to measure the amount and directionality of the change in individual scores to consensus scores.

4. Track individual examiners across several years, comparing the variation in scores compared to consensus scores to determine if experience level influences scoring stability.
5. Analyze examiners within teams to determine distance from consensus and isolate influencing factors. For example, examine whether or not one examiner scored closest to the consensus score, whether a small group scored closest to consensus score, or a majority scored closest to consensus score. Outcomes would enable researchers to determine of one person, a small group, or the majority had the most influence and under what conditions.

Summary

This study was designed to add to existing research on third-party assessment by looking at the ability of third-party examiners to assess whether or not organizations successfully implemented continuous improvement strategies based on the Criteria of the Texas Award for Performance Excellence. In general, examiners did not score consistently with the consensus score. In the future, additional data needs to be collected in a way that would allow for repetition of individual examiners' scores or organizational assessments so that variation across examiners or across sectors can be measured. As more is learned about what causes variation or scoring instability, Quality Texas will be able to refine examiner training, making it more efficient and

effective. As a result of improved training, examiners, the scoring process and the applicant organizations themselves will benefit.

REFERENCES

About Quality Texas. (n.d.). Retrieved August 27, 2004, from http://www.texas-quality.org

APQC's Knowledge Sharing Network. (n.d.). Retrieved January 30, 2006 from http://www.apqc.org

Auerbach, C. F., & Silverstein, L. B. (2003). *Qualitative data: An introduction to coding and analysis*. New York: New York University Press.

Bell, R., & Keys, B. (1998). A conversation with Curt W. Reinmann on the background and future of the Baldrige award. *Organizational Dynamics, 26*(4), 51-62.

Berquist, T. M. (1996). The backbone of state quality awards: The examiners. *Journal for Quality and Participation, 19*(4), 78-84.

Coleman, G. D., (1996). Estimating the impact of third-party evaluator training and characteristics on the scoring of written organizational self-assessments. *Digital Dissertations*, (UMI No. 9638599).

Coleman, G. D., Koelling, C. P., & Geller, E. S. (2001). Training and scoring accuracy of organizational self-assessments. *The International Journal of Quality & Reliability Management, 18*(5), 512-527.

Coleman, G. D., Van Aken, E. M., & Shen, J. (2002). Estimating interrater reliability of examiner scoring for a state quality award. *Quality Management Journal 9*(4), 39-58.

Conti, T. (1994). Time of critical review on quality self-assessment. *The Use of Quality Award Criteria and Models of Self-assessment Purposes: Proceedings of the First European Forum on Quality Self-assessment*, (pp. 169-180). Torino, Italy: European Organization for Quality.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: SAGE Publications.

Crosby, P. B. (1979). *Quality is free: The art of making quality certain*. New York: McGraw-Hill.

DeBaylo, P. W. (1999). Ten reasons why the Baldrige model works. *The Journal for Quality and Participation, 22*(1), 24-28.

DeCarlo, N. J., & Sterett, W. K. (1990). History of the Malcolm Baldrige National Quality Award. *Quality Progress, 23*(3), 21-27.

Deming, W. E. (1986). *Out of crisis*. Cambridge, MA: Massachusetts Institute of Technology.

Dubin, R. (1978). *Theory building*. New York: Free Press.

Emory, C. W., & Cooper, D. R. (1991). *Business research methods. (4th ed.).* Homewood, IL: Richard D. Irwin, Inc.

Evans, J. R., & Jack, E. P. (2003). Validating key results linkages in the Baldrige performance excellence model. *The Quality Management Journal*, *10*(2), 7-24.

Evans, J. R., & Lindsay, W. M. (1999). *The management and control of quality*. Cincinnati, OH: South-Western.

Feigenbaum, A. V. (1983). *Total quality control*. New York: McGraw-Hill.

Fuchs, E., & Stuntebeck, S. H. (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.

Garvin, D. A. (1988). *Managing quality: The strategic and competitive edge*. New York: The Free Press.

Godfrey, A. B., & Myers, D. H. (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.

Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decisions Processes, 50,* 300-323.

Hoyer, R. W., & Hoyer, B. Y. (2001). What is quality? *Quality Progress, 34*(7), 52-62.

Ishikawa, K. (1985). *What is total quality control? The Japanese way*. Englewood Cliffs, NJ: Prentice-Hall.

Jernberg, B., Lindstrom, J., & Chocron, R. (1994). How smaller companies and smaller groups use and benefit from the criteria of an award. *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 (1989). *Juran on leadership for quality: An executive handbook.* New York: McGraw-Hill.

Juran, J. M. (1991). World War II and the quality movement. *Quality Progress, 24*(12), 19-24.

Juran, J. M. (1995). *Managerial breakthrough: The classic book on improving management performance*. New York: McGraw-Hill.

Juran, J. M. (1997). Early SQC: A historical supplement. *Quality Progress, 30*(9), 73-82.

Keinath, B. J., & Gorski, B. A. (1999). An empirical study of the Minnesota Quality Award evaluation process. *Quality Management Journal, 6*(1), 29-39.

Lindsay, W. M., & Petrick, J. A. (1997). *Total quality and organization development.* Delray Beach, FL: St. Lucie Press.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley Publishing Company.

Malcolm Baldrige National Quality Award. (1998). Ten years of business excellence for America. *National Institute for Standards and Technology*. Retrieved February 13, 2005, from http://www.nist.gov

Marchese, T. (1991). TQM reaches the academy. *AAHE Bulletin, 44*, 3-9.

Martell, R. F., & Borg, M. R. (1993). A comparison of the behavioral rating accuracy of groups and individuals. *Journal of Applied Psychology, 78*(1), 43-50.

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.

McIntyre, R. M., Smith, D. E., & Hasssett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*(1), 147-156.

Miller, D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement (6th ed.).* Thousand Oaks, CA: SAGE Publications.

National Institute for Standards and Technology (NIST). (1998). Retrieved August 27, 2004, from http://www.nist.gov

Network for Excellence. (n.d.). Retrieved February 13, 2005 from http://www.networkforexcellence.org/cgi-bin/dbman/db.cgi?db=program&uid=default&ProgramName=*&view_records=1&sb=6&so=ascend

Pryor, M. (1998). *Strategic quality management*. Houston, TX: Dame Publications, Inc.

Quality Digest. (2005). Retrieved January 30, 2006, from http://www.qualitydigest.com

Quality Texas Foundation. (2005). Texas Award for Performance Excellence. (n.d.). *Criteria for performance excellence*. Retrieved November 7, 2005, from http://www.texas-quality.org

Robson, C. (1993). *Real-world research: A resource for social scientists and practitioner-researchers*. Oxford, UK: Blackwell Publishers Ltd.

Shewart, W. A. (1931). *Economic control of quality of manufactured product*. New York: Van Nostrand.

Sienknecht, T. (1999). *An empirical analysis of rating effectiveness for a state quality award*. Unpublished master's thesis, Virginia Tech, Blacksburg, VA.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*(1), 22-40.

Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*(6), 994-1003.

Tindale, R. S. (1989). Group vs. individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes, 44*, 454-473.

United Nations Economic Commission for Europe. (2004). Retrieved January 30, 2006 from http://www.unece.org

Vokurka, R. J. (2001). The Baldrige at 14. *Journal of Quality and Participation, 24*(2), 13-20.

Vokurka, R. J., Stading, G. L., & Brazeal, J. (2000). A comparative analysis of national and regional quality awards. *Quality Progress, 33*(8), 41-49.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521-551.

APPENDIX A

POST HOC TEST RESULTS FOR OVERALL ITEM DEVIATION SCORES FOR LEVEL OF EXAMINER EXPERIENCE

Post Hoc Test for Item 1.1			
		Subset	
Experience Year	N	1	2
New	119	-12.0672	
Return	76		-3.2355
Senior	54		-.7352
Sig.		1.000	.378

Post Hoc Test for Item 1.2		
Experience Year	N	Subset
		1
New	119	-10.3017
Return	76	-5.1842
Senior	54	-4.3130
Sig.		.050

Post Hoc Test for Item 2.1			
		Subset	
Experience Year	N	1	2
New	119	-13.2092	
Return	76	-8.5921	-8.5921
Senior	54		-3.3870
Sig.		.075	.098

Post Hoc Test for Item 2.2			
		Subset	
Experience Year	N	1	2
New	119	-14.7626	
Return	76		-6.1664
Senior	54		-2.5407
Sig.		1.000	.242

Post Hoc Test for Item 3.1			
		Subset	
Experience Year	N	1	2
New	119	-13.0618	
Return	76		-6.5033
Senior	54		-5.6037
Sig.		1.000	.774

Post Hoc Test for Item 3.2			
		Subset	
Experience Year	N	1	2
New	119	-12.5660	
Return	76		-5.1493
Senior	54		-3.4981
Sig.		1.000	.567

Post Hoc Test for Item 4.1			
		Subset	
Experience Year	N	1	2
New	119	-10.4063	
Return	76	-6.0717	
Senior	54		-.1519
Sig.		.078	1.000

Post Hoc Test for Item 4.2			
		Subset	
Experience Year	N	1	2
New	119	-12.5479	
Return	76		-6.6000
Senior	54		-3.3074
Sig.		1.000	.289

Post Hoc Test for Item 5.1			
		Subset	
Experience Year	N	1	2
New	119	-8.5903	
Return	76	-4.8836	-4.8836
Senior	54		-2.4111
Sig.		.108	.376

Post Hoc Test for Item 5.2			
		Subset	
Experience Year	N	1	2
New	119	-8.4538	
Return	76		-3.6711
Senior	54		.3333
Sig.		1.000	.155

Post Hoc Test for Item 5.3			
		Subset	
Experience Year	N	1	2
New	119	-11.1723	
Return	76		–2.8632
Senior	54		-.5611
Sig.		1.000	.435

Post Hoc Test for Item 6.1		
		Subset
Experience Year	N	1
New	119	-10.8504
Return	76	-6.3974
Senior	54	-6.1074
Sig.		.125

Post Hoc Test for Item 6.2			
		Subset	
Experience Year	N	1	2
New	119	-13.9979	
Return	76		–7.6414
Senior	54		-4.9907
Sig.		1.000	.428

Post Hoc Test for Item 7.1			
		Subset	
Experience Year	N	1	2
New	119	-15.9139	
Return	76	-11.8257	
Senior	54		-4.4167
Sig.		.098	1.000

Post Hoc Test for Item 7.2			
		Subset	
Experience Year	N	1	2
New	119	-15.5597	
Return	76		-9.4276
Senior	54		-5.7981
Sig.		1.000	.233

Post Hoc Test for Item 7.3				
Experience Year	N	Subset		
		1	2	3
New	119	-16.2613		
Return	76		-9.6447	
Senior	54			-3.9648
Sig.		1.000	1.000	1.000

Post Hoc Test for Item 7.4			
		Subset	
Experience Year	N	1	2
New	119	-17.2277	
Return	76		-10.0934
Senior	54		-5.8222
Sig.		1.000	.164

APPENDIX B

POST HOC TEST RESULTS FOR OVERALL ITEM DEVIATION SCORES FOR SECTORS

Post Hoc Test for Item 3.1 Customer and Market Knowledge			
Sector	N	Subset	
		1	2
Public	22	-15.2273	
Manufactory	76	-11.7132	-11.7132
Education	45	-11.4000	-11.4000
Health Care	28	-8.3929	-8.3929
Small Organization	40	-8.2000	-8.2000
Service	39		-2.0513
Sig.		.556	.070

Post Hoc Test for Item 5.1 – Work Systems			
		Subset	
Sector	N	1	2
Public	22	-10.4545	
Small Organization	40	-8.7500	
Manufactory	76	-8.3895	-8.3895
Health Care	28	-5.8929	-5.8929
Education	45	-3.1333	-3.1333
Service	39		-.2564
Sig.		.283	.073

Post Hoc Test for Item 5.2 – Employee Learning and Motivation			
Sector	N	Subset	
		1	2
Public	22	-12.2727	
Small Organization	40	-7.9250	-7.9250
Education	45	-7.7556	-7.7556
Manufactory	76	-3.2368	-3.2368
Service	39	-2.1795	-2.1795
Health Care	28		.0000
Sig.		.062	.123

Post Hoc Test for Item 5.3 – Employee Well-Being and Satisfaction

Sector	N	Subset 1	Subset 2	Subset 3
Public	22	-15.4545		
Education	45	-10.0222		
Manufactory	76	-8.4526	-8.4526	
Small Organization	40	-4.3500	-4.3500	-4.3500
Service	39		-.2564	-.2564
Health Care	28			1.0714
Sig.		.124	.079	.585

Post Hoc Test for Item 6.1 – Value Creation Processes

Sector	N	Subset 1	Subset 2
Education	45	-15.3333	
Public	22	-12.5000	-12.5000
Small Organization	40	-9.6750	-9.6750
Service	39	-5.7692	-5.7692
Health Care	28	-5.1786	-5.1786
Manufactory	76		-5.0684
Sig.		.060	.334

Post Hoc Test for Item 6.2 – Support Processes and Operational Planning

Sector	N	Subset 1	Subset 2
Education	45	-17.7111	
Public	22	-12.5000	-12.5000
Manufactory	76	-9.5263	-9.5263
Service	39	-7.8205	-7.8205
Small Organization	40	-7.1250	-7.1250
Health Care	28		-4.2857
Sig.		.061	.579

VITA

Brandi Lyn Plunkett
3780 Copperfield Dr. #416
Bryan, TX 77802

Education

2006 Doctor of Philosophy, Educational Human Resource Development, Texas A&M University, College Station, TX

2001 Master of Science, Educational Human Resource Development, Texas A&M University, College Station, TX

1990 Bachelor of Science, Curriculum and Instruction, Texas A&M University, College Station, TX

Certifications (State Of Texas)

Professional Elementary Education (Life)
Certified Training Professional (Texas A&M University & American Society for Training and Development)

Experience

2005-Present Program Supervisor, Curriculum, Certification and Evaluation, Texas Engineering Extension Service, College Station, TX

2004-2005 Graduate Intern, Leadership Development, Institute for School-University Partnerships, Texas A&M University System, College Station, Texas

2001 - 2004 Graduate Assistant for Dr. Bryan Cole, Texas A&M University, College Station, TX

1994 - 2001 Teacher, Mason Elementary School, Leander ISD, Cedar Park, TX

This dissertation was typed and prepared by Bill A. Ashworth, Jr.