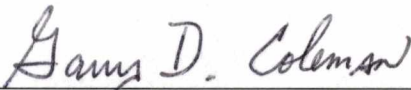


To the Graduate Council:

I am submitting herewith a thesis written by Jianming Shen entitled "An Empirical Analysis of Rating Effectiveness for a State Productivity and Quality Award." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

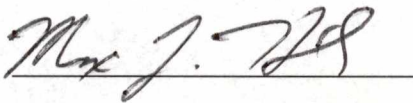


Garry D. Coleman, Major Professor

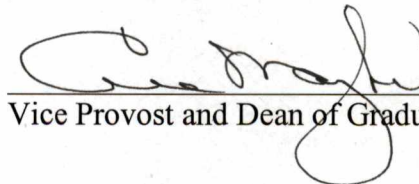
We have read this thesis
and recommend its acceptance:







Accepted for the Council:



Vice Provost and Dean of Graduate Studies

**AN EMPIRICAL ANALYSIS OF RATING EFFECTIVENESS
FOR A STATE PRODUCTIVITY AND QUALITY AWARD**

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Jianming Shen
December 2001

DEDICATION

This thesis is dedicated to my family,
who have given me the love and support.

ACKNOWLEDGEMENTS

I would like to thank Dr. Garry Coleman for his counsel and guidance during the preparation of this thesis and during my education at The University of Tennessee. I would also like to thank the members of my committee, Dr. Denise Jackson, Dr. Max Hailey and Dr. Bruce Whitehead for their guidance and support.

I owe a sincere thank-you to my academic advisor Dr. George Garrison for his guidance and support.

I would like to thank Ronald Theodore Sienknecht, Jr. for his study in rating effectiveness of the State Quality Award scoring process.

Finally, I would like to acknowledge the help of Penny Smith, Laura Horton and Mark Cross.

ABSTRACT

This study measured the rating effectiveness (interrater reliability, halo error, leniency or severity, range restriction) for each sector (Private Manufacturing Sector, Private Service Sector, Public Local Sector, and Public State and Federal Sector) of a State Quality Award using examiner scores from the 1999 award process. Results were compared with the rating effectiveness of the 1998 State Quality Award examiner scores. Interrater reliability for the Private Manufacturing Sector and range restriction for all sectors were recalculated for the 1998 State Quality Award examiner scores.

The Private Manufacturing Sector's scores exhibited low to moderate interrater reliability in both 1998 and 1999. Both the Private Service Sector's and the Public Local Sector's scores exhibited low interrater reliability in 1999, which is lower than the 1998's low to moderate interrater reliability. The Public State and Federal Sector's scores exhibited low interrater reliability in both 1998 and 1999. Halo error was present for all examiners (both years) whose halo error could be measured by the determinant of the correlation coefficient matrix. Scores from most of the dimensions in each sector showed severity in the 1999 process, while scores from most of the dimensions in each sector showed leniency in the 1998 scoring process. Scores from only a few of the dimensions in each sector showed range restriction for both the 1998 and 1999 scoring processes.

Due to the small sample size, caution should be taken in generalizing the results of this study. The small sample is a result of using field data from an actual Award rather than a controlled laboratory setting.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION AND SCOPE OF RESEARCH	1
1.1 Introduction.....	1
1.1.1 Definitions of Terms.....	3
1.1.2 The State Quality Award	4
1.1.2.1 The Award Levels.....	6
1.1.2.2 The Award Criteria	7
1.1.2.3 The Award Rating Scale.....	7
1.1.2.4 The Award Scoring Process.....	12
1.1.3 Need for Research.....	15
1.2 Problem Approach	15
1.3 Research Objectives.....	16
1.4 Research Questions	17
1.4.1 Research Questions for Interrater Reliability.....	17
1.4.2 Research Questions for Halo Error	18
1.4.3 Research Questions for Leniency or Severity.....	18
1.4.4 Research Questions for Range Restriction	18
1.5 Applications and Delimitation of the Research	18
II. LITERATURE REVIEW	20
2.1 Introduction of Rating Effectiveness	20
2.2 Methodological Issues	23
2.3 Interrater Reliability.....	24
2.4 Halo Error	27
2.5 Leniency or Severity	30
2.6 Central Tendency and Restriction of Range	32
2.7 Rating Effectiveness of Quality Awards.....	34
III. RESEARCH METHODOLOGY.....	37
3.1 Materials and Equipment	37
3.2 Data Collection	37
3.3 Data Analysis Method.....	38
3.3.1 Interrater Reliability.....	40
3.3.2 Halo Error	40
3.3.3 Leniency or Severity	41
3.3.4 Range Restriction.....	42
IV. RESULTS AND ANALYSIS.....	44
4.1 Rating Effectiveness Results for the 1998 SQA Scoring Process	44
4.1.1 Interrater Reliability for the 1998 SQA Scoring Process.....	44
4.1.2 Halo Error for the 1998 SQA Scoring Process	49

CHAPTER	PAGE
4.1.3 Leniency or Severity for the 1998 SQA Scoring Process.....	52
4.1.4 Range Restriction for the 1998 SQA Scoring Process.....	54
4.1.5 Overall Conclusion for the Rating Effectiveness of 1998 SQA Scoring Process.....	59
4.2 Interrater Reliability Results and Analysis for the 1999 SQA Scoring Process	59
4.2.1 Interrater Reliability Results and Analysis for the 1999 SQA Private Manufacturing Sector	60
4.2.2 Interrater Reliability Results and Analysis for the 1999 SQA Private Service Sector	66
4.2.3 Interrater Reliability Results and Analysis for the 1999 SQA Public Local Sector	72
4.2.4 Interrater Reliability Results and Analysis for the 1999 SQA Public State and Federal Sector	77
4.3 Halo Error Results and Analysis for the 1999 SQA Scoring Process	82
4.4 Leniency or Severity Results and Analysis for the 1999 SQA Scoring Process	84
4.4.1 Leniency or Severity Results and Analysis for the 1999 SQA Private Manufacturing Sector.....	84
4.4.2 Leniency or Severity Results and Analysis for the 1999 SQA Private Service Sector	86
4.4.3 Leniency or Severity Results and Analysis for the 1999 SQA Public Local Sector	86
4.4.4 Leniency or Severity Results and Analysis for the 1999 SQA Public State and Federal Sector.....	86
4.5 Range Restriction Results and Analysis for the 1999 SQA Scoring Process	87
4.5.1 Range Restriction Results and Analysis for the 1999 SQA Private Manufacturing Sector.....	88
4.5.2 Range Restriction Results and Analysis for the 1999 SQA Private Service Sector	90
4.5.3 Range Restriction Results and Analysis for the 1999 SQA Public Local Sector	90
4.5.4 Range Restriction Results and Analysis for the 1999 SQA Public State and Federal Sector.....	90
V. CONCLUSIONS.....	92
REFERENCES	99
APPENDICES	106
APPENDIX A 1999 SQA RAW SCORES	107

CHAPTER	PAGE
APPENDIX B INTERRATER RELIABILITY CALCULATION.....	110
APPENDIX C HALO ERROR CALCULATION	138
VITA.....	142

LIST OF TABLES

TABLE	PAGE
1.1. Definition of Rating, Rater, Ratee	4
1.2. Rating Effectiveness Defined	5
1.3. 1999 SQA Criteria (State Quality Award, 1999).....	8
1.4. Comparison of 1999 SQA Criteria with 1999 MBNQA Criteria	10
1.5. 1999 SQA Scoring Guidelines for Approach/Deployment	13
1.6. 1999 SQA Scoring Guidelines for Results	14
1.7. Timeline for the 1999 State Quality Award.....	15
1.8. Comparison of 1999 SQA Criteria with 1998 SQA Criteria	16
2.1. Rater X Ratee Matrix for Dimension d.....	23
3.1. Number of Raters and Ratees in Each Sector	38
3.2. How the Research Questions Will Be Answered	39
3.3. Data Structure for Interrater Reliability for Dimension d.....	40
3.4. Data Structure for Halo Error for Rater j.....	41
3.5. Data Structure for Leniency or Severity for Dimension d.....	42
3.6. Data Structure for Range Restriction for Dimension d.....	43
4.1. ICC(2,1) for the 1998 SQA Private Manufacturing Sector (Sienknecht, 1999)	45
4.2. Recalculation of the Interrater Reliability for the 1998 SQA Private Manufacturing Sector	46
4.3. Recalculation of the ICC(2,1) for the 1998 SQA Private Manufacturing Sector.....	47

TABLE	PAGE
4.4. ICC(2,1) for the 1998 SQA Private Service Sector	47
4.5. ICC(2,1) for the 1998 SQA Public Local Sector	48
4.6. ICC(2,1) for the 1998 SQA Public State and Federal Sector	49
4.7. Halo Error for the 1998 SQA Private Manufacturing Sector (Sienknecht, 1999)	50
4.8. Halo Error for the 1998 SQA Private Service Sector (Sienknecht, 1999)	51
4.9. Halo Error for the 1998 SQA Public Local Sector (Sienknecht, 1999)	51
4.10. Halo Error for the 1998 Public State and Federal Sector (Sienknecht, 1999)	51
4.11. Rater Main Effect R-squared Values for all 1998 Sectors (Sienknecht, 1999)	52
4.12. Mean Ratings for all 1998 Sectors (Sienknecht, 1999)	53
4.13. Variance Analysis for all 1998 Sectors (Sienknecht, 1999)	55
4.14. Range Restriction Results for all 1998 Sectors (Sienknecht, 1999)	55
4.15. Recalculated Range Restriction Results for the 1998 Private Manufacturing Sector	56
4.16. Recalculated Range Restriction Results for the 1998 Private Service Sector	57
4.17. Recalculated Range Restriction Results for the 1998 Public Local Sector	57
4.18. Recalculated Range Restriction Results for the 1998 Public State and Federal Sector	58
4.19. Recalculated Range Restriction Results for all 1998 Sectors	58
4.20. 1999 SQA ICC(1,1) Results for all Sectors	60
4.21. ICC(1,1) for the 1999 SQA Private Manufacturing Sector	61
4.22. Interquartile Ranges for the 1999 SQA Private Manufacturing Sector	61
4.23. ICC(1,1) for the 1999 SQA Private Service Sector	66

TABLE	PAGE
4.24. Interquartile Ranges for the 1999 SQA Private Service Sector	66
4.25. ICC(1,1) for the 1999 SQA Public Local Sector	71
4.26. Interquartile Ranges for the Public Local Sector	72
4.27. ICC(1,1) for the 1999 SQA Public State and Federal Sector	77
4.28. Interquartile Ranges for the 1999 SQA Public State and Federal Sector	77
4.29. Halo Error for Examiners in the 1999 SQA Private Manufacturing Sector	83
4.30. Halo Error for Examiners in the 1999 SQA Public Local Sector	83
4.31. Mean Dimension Rating Analysis for the 1999 SQA Private Manufacturing Sector.....	85
4.32. Mean Dimension Rating Analysis for the 1999 SQA Private Service Sector	85
4.33. Mean Dimension Rating Analysis for the 1999 SQA Public Local Sector	86
4.34. Mean Dimension Rating Analysis for the 1999 SQA Public State and Federal Sector	87
4.35. Range Restriction Results for the 1999 SQA Private Manufacturing Sector	89
4.36. Range Restriction Results for the 1999 SQA Private Service Sector	89
4.37. Range Restriction Results for the Public Local Sector	89
4.38. Range Restriction Results for the 1999 SQA Public State and Federal Sector	91

LIST OF FIGURES

FIGURE	PAGE
1.1. Ranked Benefits of Organization Self-assessment (Hankes, 1996).....	2
4.1. Boxplots for Interrater Reliability, the 1999 SQA Private Manufacturing Sector.....	62
4.2. Boxplots for Interrater Reliability, the 1999 SQA Private Service Sector	67
4.3. Boxplots for Interrater Reliability, the 1999 SQA Public Local Sector	73
4.4. Boxplots for Interrater Reliability, the 1999 SQA Public State and Federal Sector	78
4.5. Histogram of Entire Data Set.....	88
5.1. Examiner 1's scores	94

CHAPTER I

INTRODUCTION AND SCOPE OF RESEARCH

This chapter introduces the research topic and defines the research scope. Section 1.1 provides background information. The research problem approach is defined in Section 1.2. Section 1.3 gives the research objectives and Section 1.4 defines the research questions. The premises and delimitation of this research are outlined in Section 1.5.

A review of existing research is discussed in Chapter 2. Chapter 3 presents the research methodology. Results and analysis are presented in Chapter 4, and Chapter 5 presents the conclusions.

1.1 Introduction

Organizations are seeking every opportunity to succeed or even survive in today's rapidly changing and competitive marketplace. Award-based assessment or self-assessment has been used by organizations to identify their weaknesses and strengths whether they are applying for a productivity and quality award or not. Based on a survey, Hanks (1996) listed the ranked benefits of organization self-assessment (see Figure 1.1). It is particularly noteworthy to observe that "winning awards" is ranked second to last and that the key perceived benefit of self-assessment is about using the process(es) to help measure current successes and target future improvement activities.

Today, there is a growing tendency to adopt the assessment approach used by the quality awards for self-assessment. The Baldrige Criteria for Performance Excellence have been significant tools used by many organizations to assess and improve their

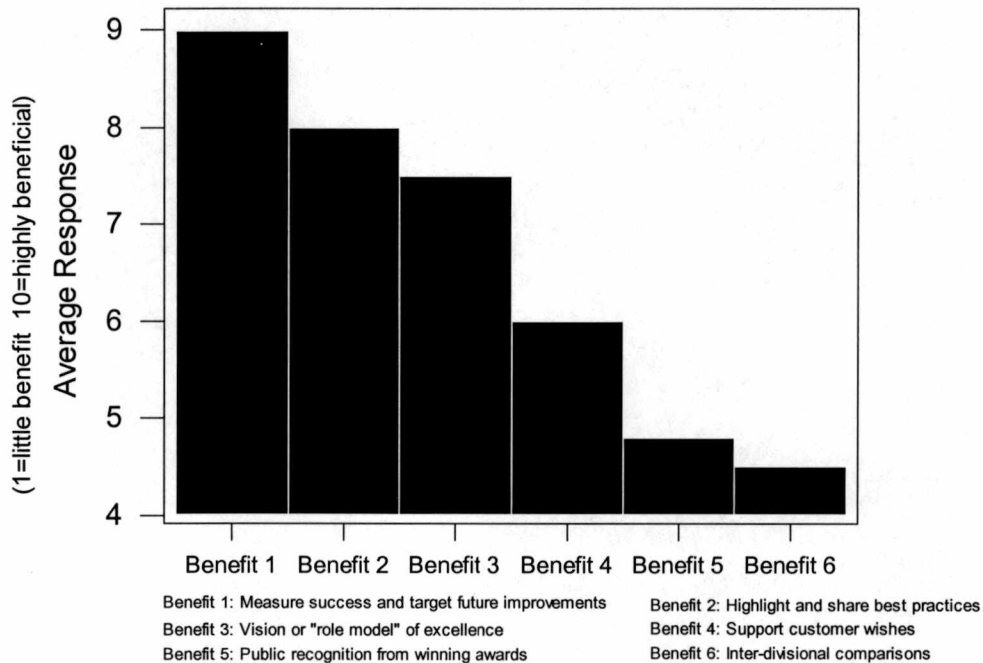


Figure 1.1. Ranked Benefits of Organization Self-assessment (Hankes, 1996)

performance. "There is no better assessment of your organization's performance available than applying for the Award (Malcolm Baldrige National Quality Award (MBNQA)). The Criteria for Performance Excellence are unique in addressing all the critical elements of your performance system." (Woolpert, in NIST, 1998, p. 2). "The application and review process for the Award (MBNQA) is the best, most cost-effective, and comprehensive business health audit you can get" (Weimerskirch, in NIST, 1998, p. 1).

Besides MBNQA, many states have their own quality awards, some of which predate the MBNQA. The web site www.qualitytoday.com/library/states-1.htm gives links to all the state quality awards, and information about these awards can be found from these

links. According to www.quality.nist.gov/98s<bl.htm, in 1998, there were 58 state and local quality award programs (44 state quality award programs, 3 state senate productivity award programs and 11 local quality award programs) and 48 states were covered. For these programs, 830 applications were received and 65,209 criteria were distributed in 1998. Many organizations use the criteria for other purposes, than actually use it to apply for the award (Bemowski and Stratton, 1995).

To the extent that quality awards are used in any form, it is important to consider the effectiveness of the instruments and processes used for evaluation. This is especially the case for publicized quality awards, which depend on effective assessments for valid results. The rating effectiveness of an assessment can be defined as the degree to which the scoring process is yielding expected outcomes (Sienknecht, 1999). Research on rating effectiveness, particularly rater error, can be and has been generalized to the scoring of quality awards (see Coleman, 1996; Coleman and Koelling, 1998; Keinath and Gorski, 1999; Sienknecht, 1999). In light of the research findings from Coleman (1996), Coleman and Koelling (1998), Keinath and Gorski (1999), and Sienknecht (1999), this research will focus on measuring the rating effectiveness of a state quality award scoring process.

1.1.1 Definitions of Terms

A number of terms are used to describe particular classifications or elements of the State Quality Award process examined in this research. For the sake of early clarification, these terms and their synonyms are listed in Table 1.1. Note also that when the term "state quality award" is used in a general sense, it is not capitalized. However, when the specific State Quality Award (SQA) studied in this research is mentioned, the term is capitalized.

Table 1.1. Definition of Rating, Rater, Ratee

Term	Synonym	Definition
Rating	Score	The numerical outcome of a single evaluation
Rater	Examiner, SQA examiner, Judge	The person who performs the evaluation
Ratee	Applicant, SQA applicant organization	The organization being evaluated

There exists a state of confusion in the literature around conceptual and operational definitions of popular rating effectiveness criteria (Saal, Downey, and Lahey, 1980, Sienknecht, 1999). Saal et al. defined four common forms of rater error: halo, leniency or severity, central tendency and restriction of range, and interrater reliability. The literature review section will give a detailed discussion of the alignment between conceptual definitions and operational definitions for these rating effectiveness criteria. For the sake of consistency in this research, the definitions listed in Table 1.2 will be used.

1.1.2 The State Quality Award

The State Quality Award (SQA) under consideration in this research is part of the U.S. Productivity and Quality Award system of state-level awards. The awards are organized into four categories or sectors (State Quality Award, 1999):

- Private Sector Manufacturing
- Private Sector Service
- Public Sector State and Federal Agencies
- Public Sector Local Agencies

Table 1.2. Rating Effectiveness Defined

Rating Effectiveness Criterion	Conceptual Definition	Operational Definition
Halo Error	Assigning ratings to individuals by attending to a global impression of each ratee rather than by carefully distinguishing among levels of performance that individual ratees exhibit on different performance dimensions (Borman, 1975).	Measured by the correlations among different dimension ratings, using a single rater's scores (over ratees) for each dimension as data points (Pulakos, Schmitt and Ostroff, 1986). Higher correlations suggest less discrimination among different aspects of behavior and thus more halo.
Leniency or Severity	A tendency to assign a higher or lower rating to an individual than is warranted by that ratee's behavior (Saal and Landy, 1977).	Measured by the comparisons of grand average dimension ratings with the midpoint(s) of the rating scale(s) used (Bernardin, Alvares, and Cranny, 1976). Mean dimension ratings that exceed the midpoint are thought to reflect leniency, whereas mean ratings below the midpoint are thought to reflect severity.
Central Tendency and Restriction of Range	Central Tendency: The situations in which ratings are clustered about the midpoint of the rating scale, reflecting rater's reluctance to use either of the extreme ends of the continuum. Restriction of Range: The situations in which ratings are clustered about any point on the rating continuum, be it favorable (lenient) point, an unfavorable (severe) point, or the midpoint (central tendency) on the rating scale (Saal et al., 1980).	Measured by calculating the standard deviation of the ratings assigned to all ratees on a particular performance dimension (Borman and Dunnette, 1975). Smaller standard deviations reflect greater range restriction. Whether such range restriction represents leniency, severity, or central tendency can only be determined by examining the mean rating.
Interrater Reliability	The extent to which two or more raters independently provide similar ratings on given aspects of the same individuals' behaviors (Saal et al., 1980).	Measured by intraclass correlation coefficient (Shrout and Fleiss, 1979).

1.1.2.1 The Award Levels

There is no limit or maximum number of organizations that may receive recognition. It is also not mandatory for the SQA Board to present awards in each sector. There are three levels of recognition that may be awarded in each of the four award categories (State Quality Award, 1999):

- **Medallion of Excellence:** This is the highest level of recognition presented to organizations which have demonstrated through their practices and achievements the highest level of productivity and quality excellence. Recognition is provided at this level to organizations which have been identified as model organizations. These organizations have demonstrated excellence and maturity in each criterion and have produced outstanding results.
- **Plaque for Outstanding Achievement:** Recognition is provided to organizations which have demonstrated through their commitment and practice of quality principles significant process in building sound processes. Organizations recognized at this level clearly demonstrate results directly attributable to a systematic, well-deployed approach.
- **Certificate for Significant Achievement:** Recognition is provided to organizations which have demonstrated, through their commitment and implementation of quality management principles, notable progress in building sound processes. These organizations have documented a solid approach and are implementing plans and procedures.

In addition to these three levels of achievement, the SQA gives the Award for Continuing Excellence (ACE) to past recipients of the SQA Medallion that have

demonstrated “sustained exemplary performance in quality and productivity improvement” three years after their initial award:

- **Award for Continuing Excellence:** Recognition is provided for past medallion recipients demonstrating sustained exemplary performance in quality and productivity. An organization is eligible to apply for the ACE three years after receiving a medallion.

1.1.2.2 The Award Criteria

The State Quality Award applicants are assessed on the seven criteria described in Table 1.3 (State Quality Award, 1999). These criteria were modified to a “condensed Baldrige” format. Table 1.4 gives a comparison of the 1999 SQA criteria with 1999 MBNQA criteria. It seems there are little differences between these two criteria.

1.1.2.3 The Award Rating Scale

The State Quality Award scoring process uses two sets of rating scales (scoring guidelines). One set of scoring guidelines pertains to the organization’s approach and deployment, and the second set to the organization’s results (State Quality Award, 1999).

Approach refers to how the applicant addresses the item requirements – the method(s) used. The factors used to evaluate approaches include the following:

- Appropriateness of the methods to the requirements.
- Effectiveness of use of methods. Degree to which the approach is systematic, integrated, and consistently applied; embodies evaluation/improvement cycles; is based upon data and information that are objective and reliable.
- Evidence of innovation. This includes significant and effective adaptation of approaches used in other types of applications or business.

Table 1.3. 1999 SQA Criteria (State Quality Award, 1999)

Criteria (point value of 1000 total)	Description
1 Leadership (140 pts) 1.1 Organizational Leadership (100 pts) 1.2 Public Responsibility and Citizenship (40 pts)	This category examines how an organization's senior leaders address values and performance expectations, as well as a focus on customers and other stakeholders, empowerment, innovation, learning, and organizational directions. Also examined is how an organization addresses its responsibilities to the public and supports its key communities. How do senior leaders guide an organization and review organizational performance? How does an organization address its responsibilities to the public and how does an organization practice good citizenship?
2 Strategic Planning (100 pts) 2.1 Strategic Development (50 pts) 2.2 Strategy Deployment (50 pts)	This category examines an organization's strategy development process, including how an organization deploys strategic objectives and action plans. An organization's strategy development process to strengthen organizational performance and competitive position. An organization's strategy deployment process.
3 Customer and Market Focus (110 pts) 3.1 Customer and Market Knowledge (50 pts) 3.2 Customer Satisfaction and Relationships (60 pts)	This category examines how an organization determines requirements, expectations, and preferences of customers and markets. Also examined is how an organization builds relationships with customers and determines their satisfaction. How does an organization determine short- and longer-term requirements, expectations, and preferences of customers and markets to ensure the relevance of current products/services? How does an organization determine the satisfaction of customers and build relationships to retain current business and to develop new opportunities?

Table 1.3. Continued

4 Information and Analysis (80 pts)	This category examines an organization's performance measurement system and how an organization analyzes performance data and information.
4.1 Measurement of Organizational Performance (80 pts)	How does an organization provide effective performance measurement systems for understanding, aligning, and improving performance at all levels and in all parts of an organization?
5 Human Resource Focus (110 pts)	This category examines how an organization enables employees to develop and utilize their full potential, aligned with the organization's objectives. Also examined are an organization's efforts to build and maintain a work environment and an employees support climate conducive to performance excellence, full participation, and personal and organizational growth.
5.1 Work Systems (40 pts)	How does an organization's work and job design, compensation, career progression, and related work force practices enable employees to achieve high performance in the organization?
5.2 Employee Education, Training, and Development (40 pts)	How does an organization's education and training support the achievement of the business objectives, build employee knowledge, skills, and capabilities, and contribute to improved employee performance?
5.3 Employee Well-being and Satisfaction (30 pts)	How does an organization maintain a work environment and an employee support climate that contribute to the well-being, satisfaction, and motivation of all employees?
6 Process Management (100 pts)	This category examines the key aspects of an organization's process management, including customer-focused design, product and service delivery, support, and supplier and partnering processes involving all work units.
6.1 Product and Service Processes (75 pts)	How does an organization manage its key product/service or support process design and/or delivery processes?
6.2 Supplier and Partnering Processes (25 pts)	How does an organization manage its key supplier and/or partnering interactions and processes?

Table 1.3. Continued

7 Business Results (360 pts)	This category examines an organization's performance and improvement in key business areas – customer satisfaction, financial and marketplace performance, and human resource results.
7.1 Customer Focused Results (90 pts)	An organization's customer focused results, including customer satisfaction and product and service performance results.
7.2 Financial and Market Results (90 pts)	An organization's key financial and marketplace performance results, segmented by market segments, as appropriate.
7.3 Human Resource Results (50 pts)	An organization's human resource results, including employee well-being, satisfaction, development, and work system performance.
7.4 Supplier and Partner Results (40 pts)	An organization's key supplier and partner results.
7.5 Organizational Effectiveness Results (90 pts)	An organization's key operational performance results that contribute to the achievement of organizational effectiveness.

Table 1.4. Comparison of 1999 SQA Criteria with 1999 MBNQA Criteria

SQA Criteria (point value of 1000 total)	MBNQA Criteria (point value of 1000 total)
1 Leadership (140 pts)	1 Leadership (125 pts)
1.1 Organizational Leadership (100 pts)	1.1 Organizational Leadership (85 pts)
1.2 Public Responsibility and Citizenship (40 pts)	1.2 Public Responsibility and Citizenship (40 pts)
1 Strategic Planning (100 pts)	1 Strategic Planning (85 pts)
2.1 Strategic Development (50 pts)	2.1 Strategic Development (40 pts)
2.2 Strategy Deployment (50 pts)	2.2 Strategy Deployment (45 pts)

Table 1.4. Continued

3 Customer and Market Focus (110 pts)	3 Customer and Market Focus (85 pts)
3.1 Customer and Market Knowledge (50 pts)	3.1 Customer and Market Knowledge (40 pts)
3.2 Customer Satisfaction and Relationships (60 pts)	3.2 Customer Satisfaction and Relationships (45 pts)
4 Information and Analysis (80 pts)	4 Information and Analysis (85 pts)
4.1 Measurement of Organizational Performance (80 pts)	4.1 Measurement of Organizational Performance (40 pts)
	4.2 Analysis of Organizational Performance (45 pts)
5 Human Resource Focus (110 pts)	5 Human Resource Focus (85 pts)
5.1 Work Systems (40 pts)	5.1 Work Systems (35 pts)
5.2 Employee Education, Training, and Development (40 pts)	5.2 Employee Education, Training, and Development (25 pts)
5.3 Employee Well-being and Satisfaction (30 pts)	5.3 Employee Well-being and Satisfaction (25 pts)
6 Process Management (100 pts)	6 Process Management (85 pts)
6.1 Product and Service Processes (75 pts)	6.1 Product and Service Processes (55 pts)
6.2 Supplier and Partnering Processes (25 pts)	6.2 Support Processes (15 pts)
	6.3 Supplier and Partnering Processes (15 pts)
7 Business Results (360 pts)	7 Business Results (450 pts)
7.1 Customer Focused Results (90 pts)	7.1 Customer Focused Results (115 pts)
7.2 Financial and Market Results (90 pts)	7.2 Financial and Market Results (115 pts)
7.3 Human Resource Results (50 pts)	7.3 Human Resource Results (80 pts)
7.4 Supplier and Partner Results (40 pts)	7.4 Supplier and Partner Results (25 pts)
7.5 Organizational Effectiveness Results (90 pts)	7.5 Organizational Effectiveness Results (115 pts)

Deployment refers to the extent to which the applicant's approach is applied to all requirements of the item. The factors used to evaluate deployment include the following:

- Use of the approach in addressing business and item requirements.
- Use of the approach by all appropriate work units.

Results refer to outcomes in achieving the purposes given in the item. The factors used to evaluate results include the following:

- Current performance levels.
- Performance levels relative to appropriate comparisons and/or benchmarks.
- Rate, breadth, and importance of performance improvements.
- Demonstration of sustained improvement and/or sustained high-level performance.

The 1999 SQA scoring guidelines for Approach/Deployment and Results are shown in Table 1.5 and Table 1.6 respectively (State Quality Award, 1999). Note that the Approach/Deployment scoring guidelines are used by criteria 1 through criteria 6, and the Results scoring guidelines are used only for criteria 7. It seems there is no difference between the 1999 SQA scoring guidelines and the 1999 Baldrige scoring guidelines.

1.1.2.4 The Award Scoring Process

The timeline for the State Quality Award scoring process is shown in Table 1.7 (State Quality Award, 1999). This research is concerned with conducting rating effectiveness analyses on the initial individual scoring of the written applications (This scoring happens after November 1 and before January.). However, findings from this research will have implications for the entire process.

Table 1.5. 1999 SQA Scoring Guidelines for Approach/Deployment

Score	Approach/Deployment
0%	• no systematic approach evident; anecdotal information
10% to 20%	• beginning of a systematic approach to the basic purposes of the Item • major gaps exist in deployment that would inhibit progress in achieving the basic purposes of the Item • early stages of a transition from reacting to problems to a general improvement orientation
30% to 40%	• a sound, systematic approach, responsive to the overall purposes of the Item • approach is deployed, although some areas of work units are in early stages of deployment • beginning of a systematic approach to evaluation and improvement of basic Item process
50% to 60%	• a sound, systematic approach, responsive to the overall purposes of the Item • approach is well deployed, although deployment may vary in some areas of work units • a fact-based, systematic evaluation and improvement process is in place for basic Item process • approach is aligned with basic organizational needs identified in the other Criteria Categories
70% to 80%	• a sound, systematic approach, responsive to the multiple requirements of the Item • approach is well deployed; with no significant gaps • a fact-based, systematic evaluation and improvement process and organizational learning/sharing are key management tools; clear evidence of refinement and improved integration as a result of organizational-level analysis and sharing • approach is well integrated with organizational needs identified in the other Criteria Categories
90% to 100%	• a sound, systematic approach exists which is not subject to turnover in leadership • approach is fully deployed without significant weakness or gaps in any areas or work units • a very strong, fact-based, systematic evaluation and improvement process and extensive organizational learning/sharing are key management tools; strong refinement and integration, backed by excellent organizational-level analysis and sharing • approach is fully integrated with organizational needs identified in the other Criteria Categories

Table 1.6. 1999 SQA Scoring Guidelines for Results

Score	Results
0%	no results or poor results in areas reported
10% to 20%	- some improvements and/or early good performance levels in a few areas - results not reported for many to most areas of importance to the organization's key business requirements
30% to 40%	- improvements and/or good performance levels in many areas of importance to the organization's key business requirements - early stages of developing trends and obtaining comparative information - results 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 reported for most areas of importance to the organization's key business requirements - no pattern of adverse trends and no poor performance levels 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 - business 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 very good relative performance levels - business results address most key customer, market, process, and action plan requirements
90% to 100%	- current performance is excellent in most areas of importance to the organization's key business requirements - excellent improvement trends and/or sustained excellent performance levels in most areas - evidence of industry and benchmark leadership demonstrated in many areas - business results fully address key customer, market, process, and action plan requirements

Table 1.7. Timeline for the 1999 State Quality Award

Timeframe	Event
July/August	Applications available for distribution or available on the web site.
October 1	Intent to apply due. Fax a completed copy of the application cover sheet
November 1	Application submission deadline
January	Selection of site visits by SQA Board
February	Site visits
March	Selection of finalists and final award determination by Board
April 20	Forum For Excellence and SQA Awards Celebration

1.1.3 Need for Research

The majority of rating effectiveness literature is on individual-level performance appraisals, and there is little research carefully examining the rating effectiveness of major quality awards' organizational assessments or self-assessments. Sienknecht (1999) measured the rating effectiveness (interrater reliability, halo error, leniency or severity, and range restriction) of the 1998 SQA scoring process, but since then, the SQA criteria have been modified to a "condensed Baldrige" format. The comparison of the 1999 SQA criteria with the 1998 SQA criteria is shown in Table 1.8. This study will measure the rating effectiveness (interrater reliability, halo error, leniency or severity, central tendency and restrictions of range) of the 1999 SQA scoring process, and make a comparison of the findings with those from Sienknecht's research. Since the 1999 SQA criteria are in "condensed Baldrige" format, the findings from this research may be more applicable to improving the MBNQA or other state quality award processes.

1.2 Problem Approach

This research (1) selects the proper methods for assessing rating effectiveness based on literature review, (2) employs these methods to analyze the rating effectiveness of the State Quality Award based on the 1999 award criteria and 1999 scoring data, and (3)

Table 1.8. Comparison of 1999 SQA Criteria with 1998 SQA Criteria

1999 SQA Criteria (point value of 1000 total)	1998 SQA Criteria (point value of 1000 total)
1 Leadership (140 pts)	2 Top Management Commitment and Involvement (Leadership) (140 pts)
2 Strategic Planning (100 pts)	5 Plan for Continuous Improvement (120 pts)
3 Customer and Market Focus (110 pts)	7 Customer and Supplier Involvement (120 pts)
4 Information and Analysis (80 pts)	6 Performance Measurement Process (Use of Information) (120 pts)
5 Human Resource Focus (110 pts)	3 Employee Involvement, Development, and Management of Participation (120 pts)
6 Process Management (100 pts)	4 Recognition and Rewards Systems (100 pts)
7 Business Results (360 pts)	7 Customer and Supplier Involvement (120 pts)
	8 Results Over Time (140 pts)
	1 Maturity of Effort (140 pts)

Note: for comparison purposes, the 1998 SQA criteria are not listed in order and 1998's criterion 7 is repeated.

compares the research findings with those based on the 1998 award criteria and scoring data.

1.3 Research Objectives

This research has four objectives:

1. Analyze the rating effectiveness of the SQA scoring process.
2. Propose improvements to the SQA scoring process.
3. Add to the empirical data by providing a second year of analysis for the same SQA award.
4. Improve on Sienknecht's (1999) approach of assessing the SQA rating effectiveness.

1.4 Research Questions

The effectiveness of a scoring (rating) process is effected by many factors: the rater characteristics, the ratee characteristics, the rating instrument (rating scale characteristics), the rating context (the type of organization, the purpose of rating, etc.), the rating process characteristics (administrative constraints, individual rater strategies, etc.), and the results of the rating (raw and transformed performance information, actions based on that information, etc.) (Landy and Farr, 1980). To assess the rating effectiveness of the State Quality Award, this research will evaluate the interrater reliability, halo, leniency or severity, and range restriction of the award ratings. The data used in this research are the most recent data available – 1999 SQA scoring process data. The SQA scoring process involves a Rater by Ratee by Dimension design for each sector, in which a subset of Raters (examiners) assess a subset of Ratees (applicants) on all the seven SQA criteria (Dimensions).

This research will focus on the following questions. For the sake of brevity, research questions will be numbered as Q#_.

1.4.1 Research Questions for Interrater Reliability

The following research questions address the interrater reliability:

Q1a: How much interrater reliability is present in the ratings?

Q1b: What differences in interrater reliability exist among dimensions within a given sector?

Q1c: What differences in interrater reliability exist in a given sector between 1998 scores and 1999 scores?

1.4.2 Research Questions for Halo Error

The following research questions relate to the halo error:

Q2a: How much halo error is present in the ratings?

Q2b: What differences in halo error exist among examiners within a given sector?

Q2c: What differences in halo error exist in a given sector between 1998 scores and 1999 scores?

1.4.3 Research Questions for Leniency or Severity

The following research questions relate to the leniency or severity:

Q3a: How much leniency or severity is present in the ratings?

Q3b: What differences in leniency or severity exist among dimensions within a given sector?

Q3c: What differences in leniency or severity exist in a given sector between 1998 scores and 1999 scores?

1.4.4 Research Questions for Range Restriction

The following questions address the range restriction:

Q4a: How much range restriction is present in the ratings?

Q4b: What differences in range restriction exist among dimensions within a given sector?

Q4c: What differences in range restriction exist in a given sector between 1998 scores and 1999 scores?

1.5 Applications and Delimitation of the Research

This research is concerned with rating effectiveness of organizational assessments or self-assessments based on quality awards. The problem is that rating based on the award

criteria is not an exact science and does require a heavy element of judgement. Caution should be taken to ensure the rating's effectiveness. If we know the current rating's effectiveness or ineffectiveness, we can make improvements in the future rating process.

Every year, both the raters and ratees differ. The 1999 SQA award criteria were modified to a "condensed Baldrige" format. It is infeasible to control the applicants and examiners. Without these controls, this research is inherently limited in giving general conclusions about the effects of rater's and ratee's characteristics and the criteria change on rating effectiveness. The comparison between years is also limited due the same reasons. However, this research provides empirical data where little currently exists.

CHAPTER II

LITERATURE REVIEW

This chapter discusses published research on rating effectiveness in general and quality awards in particular. Section 2.1 introduces rating effectiveness. Section 2.2 discusses methodological issues. Conceptual and operational definitions of interrater reliability, halo error, leniency or severity, and central tendency and restriction of range are given in Section 2.3, Section 2.4, Section 2.5 and Section 2.6 respectively. Section 2.7 discusses rating effectiveness of quality awards.

2.1 Introduction of Rating Effectiveness

Rating is the most ubiquitous method of performance appraisal (Landy and Farr, 1980), and it is used by many organizations for assessments or self-assessments.

Organizational assessments involving rating scales are popular and many quality awards use rating scales to score organizations' applications for the awards. Applicants of the SQA are evaluated and scored by a team of examiners based on the rating scale of each criterion and a total score is given. The final score is obtained by calculating the average total score of the examiner team. The award board uses the average total score, as well as the average score of each criterion, to assist them in deciding which organizations will be selected for site visits.

Rating is also important in organizations' self-assessments. "Self-assessment sets out to discover the causes of the weakness revealed by results, by analyzing processes and systemic factors. Its main purpose is to identify inadequate capabilities and the necessary improving action. Weights (rating, scoring) could be eliminated altogether in

improvement-oriented self-assessment, although in some ways they can be useful, if approached with caution. Where possible, however, the company should try to make an overall quantitative assessment of each category – give it a score, in other words. This is useful for an overall appraisal of each category, and essential if the categories are to be combined taking their relative importance into account. It is also useful as a general indication when the company is considering benchmarking with other organizations.” (Conti, 1997, p. 54).

“Organizational self-assessment does not mean that the organization performs the assessment entirely with its own resources. Indeed, organizations are rarely able to identify their weakness on their own. The organization will employ all the external and internal resources it considers necessary, but will not outsource responsibility for a process that is so closely linked to its key strategic decisions. Third-party assessments, such as quality award assessments, are more reliable than pure self-assessments (performed by the company itself, not by the third party) because the organization is scrutinized by an independent observer.” (Conti, 1997, p. 1).

“Little is known about the measurement properties of the organizational self-assessment process, particularly the consistency of third party evaluator scoring that provides the validation for self-assessments. A critical step toward improving the self-assessment process is understanding its properties and designing strategies for improvement based on this understanding.” (Coleman and Koelling, 1998, p. 31).

There is little in the literature on rating effectiveness of the organizational assessments or self-assessments, but the measurement of performance has occupied the attention of psychologists for many years. There is much literature on the rating

effectiveness of individual performance appraisals. Although these studies apply to personnel assessment, the methods and results can be generalized to organizational assessment as well. The most ubiquitous method of performance appraisal is rating. Unfortunately, rating has been shown to be prone to various types of systematic and random error. "The literature in the area of performance rating is fragmented. Some do research on different rating formats, whereas others examine characteristics of raters and ratees. This insularity has tended to obscure the fact that performance rating is best thought of as a system comprising many different classes of variables. The system might include the following classes of variables: (a) the roles (rater and ratee), (b) the vehicle (the rating instrument), (c) the rating context (the type of organization, the purpose for rating, etc.), (d) the rating process (administrative constraints, individual rater strategies, etc.), and (e) the results of the rating (raw and transformed performance information, actions based on that information, etc.)." (Landy and Farr 1980, p. 73). All of these variables might have effects on the rating effectiveness.

Many studies of performance appraisal focused on issues related to rating error – inadequacies of one sort or another in the ratings (Saal, Downey, and Lahey, 1980). However, there is less than unanimous agreement regarding conceptual and operational definitions (statistical indices) for the measures of rating effectiveness. This research will focus on the four most common measures – interrater reliability, halo error, leniency or severity, and central tendency and restriction of range. The following sections give a discussion of these four measures, but first a discussion of methodological issues is needed.

2.2 Methodological Issues

In the literature, there are a variety of different types of research designs, or data collection procedures, used by different researchers (Saal, et al., 1980). The ideal data collection strategy is a complete Rater X Ratee X Dimension matrix, in which all of the raters evaluate all of the ratees on all of the behavior dimensions. Unfortunately, this strategy for collecting rating data is seldom feasible, especially outside of a rigorously controlled laboratory situation. The most commonly reported design is a partial Rater X Ratee X Dimension matrix, in which some of the raters evaluate some of the ratees on all of the dimensions (Borman and Dunnette, 1975; Casio and Valenzi, 1977).

Table 2.1 consists of a Rater X Ratee matrix of hypothetical ratings on behavior dimension d. The following operational definitions of the four measures of rating effectiveness are based on this table.

Table 2.1. Rater X Ratee Matrix for Dimension d

Ratee	Rater					Mean	Variance
	1	2	3	...	j		
1	X_{11d}	X_{12d}	X_{13d}	...	X_{1jd}	$M_{1.d}$	$S^2_{X_{1.d}}$
2	X_{21d}	X_{22d}	X_{23d}	...	X_{2jd}	$M_{2.d}$	$S^2_{X_{2.d}}$
3	X_{31d}	X_{32d}	X_{33d}	...	X_{3jd}	$M_{3.d}$	$S^2_{X_{3.d}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	X_{i1d}	X_{i2d}	X_{i3d}	...	X_{ijd}	$M_{i.d}$	$S^2_{X_{i.d}}$
Mean	$M_{.1d}$	$M_{.2d}$	$M_{.3d}$	...	$M_{.jd}$	$M_{.d}$	$S^2_{M_{.jd}}$
Variance	$S^2_{X_{.1d}}$	$S^2_{X_{.2d}}$	$S^2_{X_{.3d}}$	...	$S^2_{X_{.jd}}$	$S^2_{M_{i.d}}$	$S^2_{X_{..d}}$

Note: Subscript dot indicates summing over that factor.

2.3 Interrater Reliability

Interrater reliability is the extent to which two or more raters independently provide similar ratings on given aspects of the same individual's behaviors (Saal, et al., 1980).

Based on literature review, Saal et al. (1980) found the following five different operational definitions of interrater reliability:

- (1) Measured by the variance (or standard deviation) of the ratings assigned to a particular ratee by several raters for a given dimension (Bernardin, 1977):

$$S^2_{x_{i,d}} \quad (2.1)$$

Smaller standard deviations for each of the individual behavior dimensions reflect greater interrater reliability. Formula (2.1) will be used to measure the interrater reliability for the 1999 SQA scoring process.

- (2) Measured by the correlations between pairs of raters who evaluate the same ratee on identical dimensions (Bernardin, Alvares, and Cranny, 1976):

$$r_{x_{ij}, x_{ij}} \quad (2.2)$$

for all possible combinations of raters, yielding a $j \times j$ correlation matrix (Average correlations, based on Fisher's r to z transformation, over all possible rater combinations are often reported). Larger correlations suggest greater interrater reliability.

- (3) Measured by the intraclass correlation coefficients (ICC) that are the most frequently used reliability coefficients. Six methods for estimating coefficients of interrater reliability were critiqued by Berk (1979). Five of the reliability statistics produced precise estimates, yet possessed similar limitations. Only the intraclass correlation-

generalizability theory approach seemed to offer the precision, comprehensiveness, and flexibility required to deal with the complexity of reliability assessment. Shrout and Fleiss (1979) gave the guidelines for choosing among the different forms of the intraclass correlation coefficients. In a typical interrater reliability study, each of a random sample of n ratees is rated independently by k raters. Three different cases of this kind of study can be defined (Shrout and Fleiss, 1979):

- Case 1: Each ratee is rated by a different set of k raters, who are randomly selected from a larger population of raters.

The reliability of a set of raters' individual ratings:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS} \quad (2.3)$$

BMS – between mean square in a Rater X Ratee ANOVA

WMS – within mean square in a Rater X Ratee ANOVA

For the 1999 SQA rating process in this research, each applicant organization was rated by a different set of raters from a larger rater population. Although the number of raters rating each applicant organization is not equal, formula (2.3) is still the proper one to be used in this study to measure the interrater reliability (a small part of the data points can be randomly kicked off to make equal the number of raters rating each applicant organization).

The reliability of a set of raters' mean ratings:

$$ICC(1,k) = \frac{BMS - WMS}{BMS} \quad (2.4)$$

- Case 2: A random sample of k raters is selected from a larger population, and each rater rates each ratee, that is, each rater rates n ratees altogether.

The reliability of a set of raters' individual ratings:

$$ICC(2,1) = \frac{BMS-EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n} \quad (2.5)$$

EMS – error mean square in a Rater X Ratee ANOVA

JMS – rater's mean square in a Rater X Ratee ANOVA

The reliability of a set of raters' mean ratings:

$$ICC(2,k) = \frac{BMS-EMS}{BMS+(JMS-EMS)/n} \quad (2.6)$$

- Case 3: Each ratee is rated by each of the same k raters, who are the only raters of interest.

The reliability of a set of raters' individual ratings:

$$ICC(3,1) = \frac{BMS-EMS}{BMS+(k-1)EMS} \quad (2.7)$$

The reliability of a set of raters' mean ratings:

$$ICC(3,k) = \frac{BMS-EMS}{BMS} \quad (2.8)$$

Futrell (1995) gave some examples showing how to calculate these coefficients. If the software (such as SPSS) is not available, the procedures in Futrell's paper can be followed to get these coefficients. The larger the ICC value, the higher the interrater reliability.

- (4) Measured by a statistically significant ratee main effect, especially one that explains a sizable proportion of the rating variance, in a Rater X Ratee X Dimension ANOVA (Friedman and Cornelius, 1976):

$$F = \frac{BMS}{EMS} \quad (2.9)$$

This approach is identical to one of the operational definitions of range restriction.

The presence of a significant ratee main effect can be interpreted as both high interrater reliability and the absence of range restriction, where as the absence of a significant ratee main effect can indicate both range restriction and low interrater reliability.

- (5) Measured by the absence of a significant of Rater X Ratee interaction in a Rater X Ratee X Dimension ANOVA (Heneman et al., 1975):

$$\text{Absence of } \frac{EMS}{WMS} \quad (2.10)$$

This approach is identical to one of the operational definitions of halo. The presence of a statistically significant Rater X Ratee interaction can be interpreted as both low interrater reliability and halo (Dickson and Tice, 1977; Heneman et al., 1975).

2.4 Halo Error

Halo error is assigning ratings to individuals by attending to a global impression of each ratee rather than by carefully distinguishing among levels of performance that individual ratees exhibit on different performance dimensions (Borman, 1975). There are four different operational definitions of the halo error (Saal, et al., 1980):

- (1) Measured by the intercorrelations between pairs of different dimension ratings, using an individual rater's scores (over ratees) for each dimension as data points (Pulakos, Schmitt and Ostroff, 1986):

$$r_{x_{jd}, x_{jd'}} \quad (2.11)$$

Yields a $d \times d$ correlation matrix, where d' is another dimension from among the d dimensions. These correlations are transformed to z scores using Fisher's r -to- z transformation and are then averaged to provide an average absolute level of the interdimension correlation for each rater.

Higher correlations suggest less discrimination among different aspects of behavior and thus more halo. Formula (2.11) was used in this research to measure the halo error in the SQA rating process. When the determinant of the correlation matrix was close to zero, this was interpreted as evidence of halo error (Sienknecht, 1999).

- (2) Measured by the results of a factor or principal-component analysis of the dimension intercorrelation matrix (Kraut, 1975; Blanz and Ghiselli, 1972). The fewer factors or principal components that emerge, the greater halo.
- (3) Measured by the variance (or standard deviation) associated with a particular rater's ratings of a particular ratee across all the performance dimensions (Bernardin and Walter, 1977):

$$S^2_{x_{ij}} \quad (2.12)$$

Less dispersion among the dimension ratings, evidenced by smaller standard deviation or variance estimates, indicates a greater halo error. The use of an average

standard deviation across rating dimensions within ratees as a measure of halo is criticized by Pulakos, Schmitt and Ostroff (1986) for those cases in which the average of the ratings computed within raters on each dimension is not equal. They stated that standardizing the data for each rater separately within each dimension prior to computing standard deviations across the dimensions corrects this problem. They recommended that each rater's average observed intercorrelation among the dimensions be used as a halo measure (see formula (2.11)).

- (4) Measured by the Rater X Ratee interaction based on a Rater X Ratee X Dimension ANOVA (Dickson and Tice, 1977):

$$F = \frac{EMS}{WMS} \quad (2.13)$$

A statistically significant Rater X Ratee interaction, especially one that explains a sizable proportion of the rating variance, has been interpreted as a sign of halo. This approach is identical to one of the operational definitions interrater reliability. A statistically significant Rater X Ratee interaction can be interpreted as low interrater reliability as well.

Although the halo error is one of the oldest and most widely known of rating errors, surprisingly little is known about its nature. Most studies of halo error are in personnel assessment, but the findings can be used in organizational assessment as well. Nisbett and Wilson's (1977) experimental research shows that global evaluations of a person can induce altered evaluations of the person's attributes, even when there is sufficient information to allow for independent assessments of them. Furthermore, the subjects (raters) were unaware of this influence of global evaluations on ratings of attributes. In

fact, the subjects (raters) actually believed that the direction of influence was opposite to the true direction.

Kozlowski, Kirsch and Chao's (1986) research shows that the degree of halo was significantly greater when raters lacked relevant job and ratee knowledge. Murphy and Bazler's (1986) research shows that memory-based behavior ratings and performance evaluations showed more halo than did ratings that were collected immediately after viewing the ratee's performance.

Halo is thought to result in intercorrelations among rated dimensions that are higher than the true levels of intercorrelation (Pulakos, Schmitt, and Ostroff, 1986). It is widely believed that halo in ratings reflects both true and illusory intercorrelation and that it would be possible to improve the quality of rating data if the illusory portion of halo could be controlled or removed (Bartlett, 1983). Murphy and Reynolds's (1988) experimental research showed that true halo (true intercorrelation) has some impact on observed halo (observed intercorrelation) but that this impact is relatively small. In particular, extreme differences in true halo were necessary to produce reliable differences in observed halo. In addition, their results suggested that halo error occurs only when the true intercorrelations are small. Walter C. Borman's (1975) study showed that rater training can significantly reduce halo, but it also lowered reliability (it reduced or obscured true intercorrelations).

2.5 Leniency or Severity

Leniency or Severity is a tendency to assign a higher or lower rating to an individual than is warranted by that ratee's behavior (Saal and Landy, 1977). There are three different operational definitions of leniency or severity (Saal, et al., 1980):

- (1) Measured by the comparisons of average dimension ratings with the midpoint(s) of the rating scale(s) used (Bernardin, Alvares, and Cranny, 1976):

$$M_{..d} \quad (2.14)$$

This is the most popular operational definition and was used in this study to measure the leniency or severity in the SQA rating process. Mean dimension ratings that exceed the midpoint are thought to reflect leniency, whereas mean ratings below the midpoint are thought to reflect severity. Care should be taken to the meaning of the rating scale midpoint of the quality awards. For Approach/Deployment Items, the 50% mark represents an approach that should meet the basic objectives addressed in the item. Scores higher than 50% should reflect refinement, continued progress in deployment, and cycles of improvement. For Results Items, the 50% mark represents clear indication of improvement trends and/or current good levels of performance in areas of the business covered in the Item (National Institute of Standards and Technology, 1997). Some examiners take the 50% point to mean excellence and maturity; covering all activities under the scope of an Item. This approach tends to compress the measurement scale, virtually eliminating scores of 60% or higher. This, in turn, tends to differentiate poorly among applicants, despite real differences. Though the 50% point reflects systems and results of organizations with functioning quality systems, it should not be taken to mean full deployment, maturity, and refinement (Tennessee Quality Award, 1998).

- (2) Measured by the rater main effect based on a Rater X Ratee X Dimension ANOVA (Friedman and Cornelius, 1976):

For each dimension: $F = \frac{JMS}{EMS}$ (2.15)

A statistically significant rater main effect, especially one that explains a sizable proportion of the rating variance, is interpreted as leniency or severity.

(3) Measured by the degree of skewness that characterizes frequency distributions of dimension ratings (Landy, Farr, Saal, and Freytag, 1976):

- (a) skewness of all $X_{ij,d}$ scores (includes every rater-ratee combination)
- (b) skewness of the $X_{i,d}$ scores (focuses on ratees' ratings across raters)
- (c) skewness of the $X_{j,d}$ scores (focuses on raters' ratings across ratees)

(2.16)

Significant negative skewness is thought to reflect leniency, whereas significant positive skewness is interpreted as severity.

2.6 Central Tendency and Restriction of Range

Central tendency is used to describe situations in which ratings are clustered about the midpoint of the rating scale, reflecting a rater's reluctance to use either of the extreme ends of the continuum. Restriction of range is used in situations in which ratings are clustered about any point on the rating continuum, be it a favorable (lenient) point, an unfavorable (severe) point, or the midpoint (central tendency) on the rating scale (Saal, et al., 1980).

There are four operational definitions of central tendency and restriction of range (Saal, et al., 1980):

- (1) Measured by the variance (or standard deviation) of the ratings assigned to all ratees on a particular performance dimension (Borman and Dunnette, 1975):

(a) $S^2_{X..d}$, which includes all X_{ijd} scores

(b) $S^2_{Mi.d}$, which focuses on ratees' average ratings across raters.

(2.17)

This is the most common definition and formula (a) will be used in this research to measure the central tendency and restriction of range during the SQA rating process.

Smaller standard deviations reflect greater range restriction. Whether such range restriction represents leniency, severity, or central tendency can only be determined by examining the mean rating.

- (2) Measured by the proximity of the mean dimension ratings to the midpoint of the scale (DeCotiis, 1977):

$$M_{..d} \quad (2.18)$$

- (3) Measured by the degree of kurtosis, or peakedness, that characterizes frequency distributions of dimension ratings (Landy, et al., 1976):

Variance of ratings assigned to all ratees on single performance dimension.

Kurtosis (of dimension ratings):

(a) kurtosis of all X_{ijd} scores (includes every rater-ratee combination)

(b) kurtosis of the $X_{i.d}$ scores (focuses on ratees' ratings across raters)

(c) kurtosis of the $X_{.jd}$ scores (focuses on rater's ratings across ratees).

(2.19)

- (4) Measured by the absence of a significant ratee main effect based on a Rater X Ratee X Dimension ANOVA (Heneman, Schwab, Huett, and Ford, 1975):

Absence of (for each dimension):

$$F = \frac{BMS}{EMS} \quad (2.20)$$

This approach is identical to one of the operational definitions of interrater reliability.

The presence of a significant ratee main effect can be interpreted as both high interrater reliability and the absence of range restriction, whereas the absence of a significant ratee main effect can indicate both range restriction and low interrater reliability.

2.7 Rating Effectiveness of Quality Awards

Rating effectiveness of quality awards and related self-assessments have attracted researchers' attentions. Coleman (1996) examined the impact of third-party evaluator training and characteristics on the scoring of written organizational self-assessments. The study showed how a common form of evaluator training, frame-of-reference (FOR) training, specifically affected the consistency and accuracy of the evaluators' scores. The training improved elevation accuracy and reduced a leniency effect, but did little to improve the agreement among evaluators or improve dimensional accuracy. The study illustrated how inherent differences in the content of the dimensions affected the dispersion of the scores.

Sienknecht (1999) measured the rating effectiveness of the 1998 SQA scoring. He found the interrater reliability was low to moderate and differed by dimension. Halo error was present for all examiners. Leniency or severity was present for one third of the cases and differed by dimension. Range restriction was present for two thirds of the cases and differed by dimension. No pattern existed for which dimension always exhibited leniency or severity or range restriction.

Rater training has generally been shown to be effective in reducing rating errors, especially if the training is extensive and allows for rater practice. Van der Wiele,

Williams, Kolb and Dale (1995) did an analysis of the scoring of the 1993 European Quality Award (EQA) case study by 175 participants who have undertaken assessor (rater) training. They found that those raters who have undertaken the EQA training more than once scored the criteria more stringently than those who were undertaking training for the first time (leniency), and also they produced less variation in the scoring (higher interrater reliability). They also found that the background (manufacturing or service) of the rater had no significant influence on the scoring. There are numerous studies about training's effect on rating effectiveness, such as Coleman and Koelling (1998), Bernardin and Pence (1980), Pulakos (1984), Zedeck and Cascio (1982), Borman (1979), Latham and Wexley (1975), Ivancevich (1979), Warmke and Billings (1979), Bernardin (1978), Bernardin and Walter (1977), Das, Smith and Yeager (1991). Although most of these studies are in personnel assessment, the findings can be applied to organizational assessment as well.

Rater experience appears to positively affect rating effectiveness, but the mechanism or mechanisms responsible (e.g., more training or experience with the rating instrument, better observation skills, better knowledge of the job requirements, etc.) was not known (Landy and Farr, 1980). Keinath and Gorski (1999) did an empirical study of the Minnesota Quality Award evaluation process and found that experience as an quality award examiner appears to make only a small difference in individual scoring of applications and the consensus process is effective in minimizing any undue influence of examiner experience.

Hahn and Dipboye's (1988) experimental research on the effects of training and information on interrater reliability showed that the amount of information affected

interrater reliability, and raters with more ratees' information presented higher interrater reliability than those with less information. The raters who received both training and full information showed the highest interrater reliability.

Rothstein (1990) examined the influence of the length of exposure to the ratee on the interrater reliability. It was found that there was a strong, nonlinear relationship between months of exposure and interrater reliability. The longer the length of exposure, the more the information about the ratee. It seems Rothstein's findings are consistent with those of Hahn and Dipboye's.

Murphy, Balzer, Lockhart, and Eisenman (1985) did experimental research on the effects of previous performance on evaluations of present performance and found significant contrast effect (in which evaluations of present performance were biased in a direction opposite to that of previous evaluations) and the contrast effect was not stronger when ratings were obtained from memory than when ratings were obtained immediately. From these findings, it seems an examiner's memory of an applicant organization's past performance has little effect on the examiner's current scoring of the same applicant organization if the same examiner is scoring the second application of the organization.

CHAPTER III

RESEARCH METHODOLOGY

This chapter describes the methodology used to answer the research questions discussed in Chapter 1. The necessary materials and equipment are discussed in Section 3.1. Section 3.2 discusses the data collection methodology. Section 3.3 describes the data analysis method.

3.1 Materials and Equipment

This research is an empirical analysis of rating effectiveness for a State Quality Award. The materials and equipment necessary for this analysis includes statistical software (SPSS and Minitab) and a personnel computer. Other materials include documentation of the SQA rating process, the 1999 SQA scores and related literature.

3.2 Data Collection

In the State Quality Award application process, each applicant organization is evaluated against the seven award criteria (Table 1.3). During the scoring, each sector uses a different set of examiners (raters) to rate different applicant organizations (ratees) in that sector on the same criteria (dimensions). This process yields a Rater X Ratee X Dimension design (Table 2.1), however, this design was only partially filled using the 1999 scoring data.

The data were provided by the SQA Board in a summary spreadsheet in the following format:

Examiner	Sector	Company	Item Scores			Category total			Category total		
			1.1	...	7.5	1	...	7	Total	1	...
1	1	1									
9	1	1	(data in points)			(data in points)			(data in percentage)		
...	...	...									

To facilitate the analysis, all scores (ratings) were converted from their original form as percentages to the decimal equivalent score between 0.000 and 1.000 (inclusive). The number of examiners (raters) and applicant organizations (ratees) within each sector is shown in Table 3.1. Not all examiners in each sector evaluated all the applicant organizations in that sector.

3.3 Data Analysis Method

The research questions stated in Chapter 1 will be answered by studying the results of the calculations using the formulas selected from literature (see Chapter 2). These results will be presented in tables and graphs. When comparisons are made, hypothesis testing will be performed if feasible. Table 3.2 presents the methods to answer the research questions.

The following sections present a detailed description of the statistical tests used to assess the four measures of rating effectiveness: interrater reliability, halo error, leniency or severity, and central tendency and restriction of range. These four measures, together, permit a full analysis of rating effectiveness.

Table 3.1. Number of Raters and Ratees in Each Sector

Sector	Number of Raters (Examiners)	Number of Ratees (Applicant Organizations)
Private Manufacturing	14	5
Private Service	6	2
Public State and Federal Agencies	10	2
Public Local Agencies	9	3
Total	39	12

Table 3.2. How the Research Questions Will Be Answered

Research Questions	Data Analyses	Unit of Analysis
Interrater Reliability		
Q1a: How much interrater reliability is present in the ratings?	Formula (2.1) and (2.3)	By sector and dimension
Q1b: What differences in interrater reliability exist between dimensions within a given sector?	Comparing results from Q1a dimension by dimension	
Q1c: What differences in interrater reliability exist in a given sector between 1998 scores and 1999 scores?	Comparing 1999 results with 1998 results	
Halo Error		
Q2a: How much halo error is present in the ratings?	Formula (2.11)	By examiner
Q2b: What differences in halo error exist between examiners within a given sector?	Comparing results from Q2a between examiners	
Q2c: What differences in halo error exist in a given sector between 1998 scores and 1999 scores?	Comparing 1999 results with 1998 results	
Leniency or Severity		
Q3a: How much leniency or severity is present in the ratings?	Formula (2.14)	By sector and dimension
Q3b: What differences in leniency or severity exist between dimensions within a given sector?	Comparing results from Q3a dimension by dimension	
Q3c: What differences in leniency or severity exist in a given sector between 1998 scores and 1999 scores?	Comparing 1999 results with 1998 results	
Central Tendency and Restriction of Range		
Q4a: How much range restriction is present in the ratings?	Formula (2.17)	By sector and dimension
Q4b: What differences in range restriction exist between dimensions within a given sector?	Comparing results from Q4a dimension by dimension	
Q4c: What differences in range restriction exist in a given sector between 1998 scores and 1999 scores?	Comparing 1999 results with 1998 results	

3.3.1 Interrater Reliability

Q1a will be addressed by using formula (2.1) and (2.3), and **Q1b** by a simple comparison of the calculated values from **Q1a** dimension by dimension. **Q1c** will be addressed by comparing results with those from Sienknecht's (1999) research on 1998 SQA scores. Because there are four sectors and each sector has seven dimensions (criteria), the outcome of the interrater reliability data will be 28 figures (based on formula (2.3)). The data structure used in calculating the interrater reliability is shown in Table 3.3.

3.3.2 Halo Error

Q2a will be addressed by the formula (2.11) and this yields a 7 X 7 correlation matrix with 21 non-redundant pairwise comparisons of dimensions for each examiner. The determinant of the correlation matrix serves as an aggregate measure of halo error for a

Table 3.3. Data Structure for Interrater Reliability for Dimension d

Ratee	Rater					Mean	Variance
	1	2	3	...	j		
1	X_{11d}	X_{12d}	X_{13d}	...	X_{1jd}	$M_{1.d}$	$S^2_{x_{1.d}}$
2	X_{21d}	X_{22d}	X_{23d}	...	X_{2jd}	$M_{2.d}$	$S^2_{x_{2.d}}$
3	X_{31d}	X_{32d}	X_{33d}	...	X_{3jd}	$M_{3.d}$	$S^2_{x_{3.d}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	X_{i1d}	X_{i2d}	X_{i3d}	...	X_{ijd}	$M_{i.d}$	$S^2_{x_{i.d}}$
Mean	$M_{.1d}$	$M_{.2d}$	$M_{.3d}$	...	$M_{.jd}$	$M_{.d}$	$S^2_{M_{.jd}}$
Variance	$S^2_{x_{.1d}}$	$S^2_{x_{.2d}}$	$S^2_{x_{.3d}}$	...	$S^2_{x_{.jd}}$	$S^2_{M_{i.d}}$	$S^2_{x_{..d}}$

given examiner. A determinant close to zero is evidence of halo error (Sienknecht, 1999).

Q2b will be addressed by simple comparison of the examiners' results from **Q2a**, and **Q2c** by comparing results with those from Sienknecht's (1999) research on 1999 SQA scores.

The data structure used for calculating the halo error for a given examiner is shown in Table 3.4, where $score_{ijd}$ is the j^{th} examiner's score for applicant organization i on dimension (criteria) d . If the number of ratees rated by an examiner was less than three, the halo error could not be measured by this method.

3.3.3 Leniency or Severity

Q3a will be addressed by the formula (2.14), and **Q3b** by simple comparison of results from **Q3a** dimension by dimension. **Q3c** will be addressed by comparing results with those from Sienknecht's (1999) research on 1999 SQA scores.

The data structure used for calculating the leniency or severity for a given dimension is shown in Table 3.5.

Table 3.4. Data Structure for Halo Error for Rater j

Ratee	Dimension				
	1	2	3	...	7
1	$score_{1j1}$	$score_{1j2}$	$score_{1j3}$	...	$score_{1j7}$
2	$score_{2j1}$	$score_{2j2}$	$score_{2j3}$	...	$score_{2j7}$
3	$score_{3j1}$	$score_{3j2}$	$score_{3j3}$	...	$score_{3j7}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
i	$score_{ij1}$	$score_{ij2}$	$score_{ij3}$	...	$score_{ij7}$

Table 3.5. Data Structure for Leniency or Severity for Dimension d

Ratee	Rater					Mean
	1	2	3	...	j	
1	X_{11d}	X_{12d}	X_{13d}	...	X_{1jd}	$M_{1.d}$
2	X_{21d}	X_{22d}	X_{23d}	...	X_{2jd}	$M_{2.d}$
3	X_{31d}	X_{32d}	X_{33d}	...	X_{3jd}	$M_{3.d}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	X_{i1d}	X_{i2d}	X_{i3d}	...	X_{ijd}	$M_{i.d}$
Mean	$M_{.1d}$	$M_{.2d}$	$M_{.3d}$	...	$M_{.jd}$	$M_{..d}$

3.3.4 Range Restriction

Q4a will be addressed by the formula (2.17), and **Q4b** by simple comparison of results from **Q3a** dimension by dimension. **Q4c** will be addressed by comparing results with those from Sienknecht's (1999) research on 1999 SQA scores.

The data structure used for calculating the leniency or severity for a given dimension is shown in Table 3.6.

Table 3.6. Data Structure for Range Restriction for Dimension d

Ratee	Rater					Mean	Variance
	1	2	3	...	j		
1	X_{11d}	X_{12d}	X_{13d}	...	X_{1jd}	$M_{1.d}$	$S^2_{X_{1.d}}$
2	X_{21d}	X_{22d}	X_{23d}	...	X_{2jd}	$M_{2.d}$	$S^2_{X_{2.d}}$
3	X_{31d}	X_{32d}	X_{33d}	...	X_{3jd}	$M_{3.d}$	$S^2_{X_{3.d}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	X_{i1d}	X_{i2d}	X_{i3d}	...	X_{ijd}	$M_{i.d}$	$S^2_{X_{i.d}}$
Mean	$M_{.1d}$	$M_{.2d}$	$M_{.3d}$	...	$M_{.jd}$	$M_{..d}$	$S^2_{M_{.jd}}$
Variance	$S^2_{X_{.1d}}$	$S^2_{X_{.2d}}$	$S^2_{X_{.3d}}$	...	$S^2_{X_{.jd}}$	$S^2_{M_{i.d}}$	$S^2_{X_{..d}}$

CHAPTER IV

RESULTS AND ANALYSIS

This chapter presents the results and analysis for each research question. The 1999 SQA raw scores are included in Appendix A. Computer calculation outputs are included in Appendices B and C. Section 4.1 summarizes Sienknecht's (1999) rating effectiveness results for the 1998 SQA scoring process. Section 4.2 contains interrater reliability results and analysis for the 1999 SQA scoring process. Section 4.3 contains halo error results and analysis for the 1999 SQA scoring process. Section 4.4 contains leniency or severity results and analysis for the 1999 SQA scoring process. Section 4.5 contains range restriction results and analysis for the 1999 SQA scoring process.

4.1 Rating Effectiveness Results for the 1998 SQA Scoring Process

The following summarizes Sienknecht's (1999) rating effectiveness (interrater reliability, halo error, leniency or severity, and range restriction) results for the 1998 SQA scoring process.

4.1.1 Interrater Reliability for the 1998 SQA Scoring Process

Intraclass correlation coefficient formula ICC(2,1) was used to measure the interrater reliability for all the four sectors.

Interrater Reliability for the 1998 SQA Private Manufacturing Sector

Interrater reliability results for the 1998 SQA Private Manufacturing Sector are listed in Table 4.1. Based on Shrout and Fleiss (1979), the significance test of ICC(2,1) against zero is an F test, using $F = \text{BMS}/\text{EMS}$. The main effect being tested is actually the between ratee effect, not the rater main effect as Sienknecht stated in the note to Table

Table 4.1. ICC(2,1) for the 1998 SQA Private Manufacturing Sector
(Sienknecht, 1999)

Dimension	ICC(2,1)	Significance
1 Maturity of Effort	0.272	0.000**
2 Top Management Commitment and Involvement (Leadership)	0.166	0.000**
3 Employee Involvement, Development, and Management of Participation	0.056	0.053*
4 Recognition and Rewards Systems	-0.006	0.552
5 Plan for Continuous Improvement	0.043	0.093*
6 Performance Measurement Process (Use of Information)	0.16	0.000**
7 Customer and Supplier Involvement	0.09	0.010**
8 Results over Time	0.187	0.000**

Note: significance levels (**significant at 0.05 level, *significant at 0.10 level) are based on Rater (sic) main effect significance, not testing the ICC(2,1) statistic against zero. Because this formulation of statistical significance is potentially misleading, the reader should not attach the traditional importance and interpretation to the statistical significance figures for ICC(2,1).

4.1. Sienknecht made mistakes in noting the rater main effect, then the between ratee effect, but his calculations of the significance are correct.

Sienknecht classified the ICC results as follows: low reliability (less than 0.3), moderate reliability (0.3~0.7), and high reliability (0.7~1.0) (He did not give reasons of this classification. For the purpose of comparing 1999 SQA rating effectiveness results with those of 1998 SQA scoring process, this classification will be used to 1999 SQA interrater reliability results.). So private manufacturing sector had low interrater reliability in 1998.

Based on the data structure, it seems ICC(1,1) should have been used for the Private Manufacturing Sector. The difference between ICC(1,1) and ICC(2,1) is as follows (Shrout and Fleiss, 1979):

ICC(1,1): Each ratee is rated by a different set of raters, who are randomly selected from a larger population of raters.

ICC(2,1): A random sample of raters is selected from a larger population of raters, and

Table 4.2. Recalculation of the Interrater Reliability
for the 1998 SQA Private Manufacturing Sector

Dimension	ICC(1,1)	Significance
1	0.4904	0.000**
2	0.3146	0.0007**
3	0.2492	0.0044**
4	-0.0331	0.6220
5	0.1373	0.0568*
6	0.324	0.0005**
7	0.1846	0.0212**
8	0.3836	0.0001**

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS}/\text{WMS}$, which is used to testing the ICC(1,1) against zero.

this set of raters rates all ratees.

The data structure shows that each applicant organization in the 1998 SQA Private Manufacturing Sector was rated by a different set of examiners. So ICC(1,1) should have been used for the 1998 SQA Private Manufacturing Sector, while ICC(2,1) should have been used for all the other sectors. The recalculation of the interrater reliability using ICC(1,1) for the 1998 SQA Private Manufacturing Sector is listed in Table 4.2.

Based on the recalculation, ICC values for all dimensions except dimension 4 were significantly above zero, while dimension 4 showed extremely low interrater reliability. So the 1998 SQA Private Manufacturing Sector actually exhibited low to moderate interrater reliability.

It seems Sienknecht made mistakes in calculating the ICC(2,1) for the Private Manufacturing Sector even though formula ICC(2,1) was used. Based on Shrout and Fleiss (1979), ICC(1,1) will, on the average, give smaller values than ICC(2,1) for the same data. The recalculation of the ICC(2,1) for the Private Manufacturing Sector is listed in Table 4.3, and these results are consistent with Shrout and Fleiss' (1979)

Table 4.3. Recalculation of the ICC(2,1)
for the 1998 SQA Private Manufacturing Sector

Dimension	ICC(2,1)	Significance
1	0.4930	0.000**
2	0.3172	0.0006**
3	0.2486	0.0052**
4	-0.0349	0.6293
5	0.1419	0.0494**
6	0.3345	0.0002**
7	0.1936	0.0136**
8	0.3853	0.0001**

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS/EMS}$, which is used to testing the ICC(2,1) against zero.

statement. His ICC(2,1) calculations for the other three sectors are correct.

Interrater Reliability for the 1998 SQA Private Service Sector

The interrater reliability results for the 1998 SQA Private Service Sector are listed in Table 4.4. As can be seen from the Table 4.4, ICC values for all the dimensions except dimension 3 were significantly above zero, while dimension 3 showed very low interrater reliability. So the 1998 SQA Private Service Sector exhibited low to moderate interrater reliability.

Table 4.4. ICC(2,1) for the 1998 SQA Private Service Sector (Sienknecht, 1999)

Dimension	ICC(2,1)	Significance
1 Maturity of Effort	0.443	0.002**
2 Top Management Commitment and Involvement (Leadership)	0.181	0.065*
3 Employee Involvement, Development, and Management of Participation	0.159	0.108
4 Recognition and Rewards Systems	0.202	0.046**
5 Plan for Continuous Improvement	0.18	0.043**
6 Performance Measurement Process (Use of Information)	0.24	0.014**
7 Customer and Supplier Involvement	0.422	0.001**
8 Results over Time	0.581	0.000**

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS/EMS}$, which is used to testing the ICC(2,1) against zero.

Table 4.5. ICC(2,1) for the 1998 SQA Public Local Sector

Dimension	ICC(2,1)	Significance
1 Maturity of Effort	0.427	0.006**
2 Top Management Commitment and Involvement (Leadership)	0.141	0.124
3 Employee Involvement, Development, and Management of Participation	0.086	0.202
4 Recognition and Rewards Systems	0.225	0.072*
5 Plan for Continuous Improvement	0.435	0.002**
6 Performance Measurement Process (Use of Information)	-0.167	0.989
7 Customer and Supplier Involvement	-0.013	0.454
8 Results over Time	0.175	0.048**

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS/EMS}$, which is used to testing the ICC(2,1) against zero.

Interrater Reliability for the 1998 SQA Public Local Sector

The interrater reliability results for the 1998 SQA Public Local Sector are listed in Table 4.5. As can be seen from Table 4.5, only half of the dimensions (dimension 1, 4, 5, and 8) had significantly above zero ICC values, while the other dimensions (dimension 2, 3, 6, and 7) showed very low interrater reliability. So the 1998 SQA Public Local Sector exhibited low to moderate interrater reliability.

Interrater Reliability for the 1998 SQA Public State and Federal Sector

The interrater reliability results for the 1998 SQA Public State and Federal Sector are listed in Table 4.6. As can be seen from Table 4.6, only one dimension (dimension 1) had significantly above zero ICC value, while all the other dimensions showed very low interrater reliability. So the 1998 SQA Public State and Federal Sector exhibited low interrater reliability.

For all four sectors, differences in interrater reliability existed among all dimensions within a given sector. Some dimensions exhibited moderate interrater reliability, while

Table 4.6. ICC(2,1) for the 1998 SQA Public State and Federal Sector

Dimension	ICC(2,1)	Significance
1 Maturity of Effort	0.192	0.069*
2 Top Management Commitment and Involvement (Leadership)	0.130	0.168
3 Employee Involvement, Development, and Management of Participation	0.082	0.204
4 Recognition and Rewards Systems	0.12	0.172
5 Plan for Continuous Improvement	0.064	0.283
6 Performance Measurement Process (Use of Information)	0.013	0.372
7 Customer and Supplier Involvement	0.155	0.142
8 Results over Time	0.012	0.362

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS/EMS}$, which is used to testing the ICC(2,1) against zero.

others exhibited low interrater reliability. No patterns were apparent across sectors.

4.1.2 Halo Error for the 1998 SQA Scoring Process

Intercorrelations among different dimension ratings were used as evidence of the presence or absence of halo error. These intercorrelations generate a dimension X dimension matrix for each rater. The determinant of this matrix serves as an aggregate indicator of halo error. The determinant is small when the correlations are, on average, high, and conversely large when the correlations are low. A determinant close to zero is evidence of halo error (Sienknecht, 1999).

Halo Error for the 1998 SQA Private Manufacturing Sector

The halo error results for the 1998 SQA Private Manufacturing Sector are listed in Table 4.7. For the 1998 SQA Private Manufacturing Sector, examiners 7, 8, 13, 15, 16 rated only one applicant. Correlation coefficients could not be calculated for these raters, so determinants for these raters did not exist. For examiner 14, scores on one dimension were constant (the examiner gave the same scores to different ratees). Therefore, the

Table 4.7. Halo Error for the 1998 SQA Private Manufacturing Sector (Sienknecht, 1999)

Examiner	Determinant of Dimension Interrelation Matrix	Evidence of Halo Error
1	5.187E-3	yes
2	0	yes
3	0	yes
4	0	yes
5	0	yes
6	0	yes
7	Does not exist	inconclusive
8	Does not exist	inconclusive
9	0	yes
10	0	yes
11	0	yes
12	0	yes
13	Does not exist	inconclusive
14	Does not exist	inconclusive
15	Does not exist	inconclusive
16	Does not exist	inconclusive

correlations for this dimension could not be calculated, and the determinant does not exist.

Halo Error for the 1998 SQA Private Service Sector

The halo error results for the 1998 SQA Private Service Sector are listed in Table 4.8.

Halo Error for the 1998 SQA Public Local Sector

The halo error results for the 1998 SQA Public Local Sector are listed in Table 4.9.

Halo Error for the 1998 SQA Public State and Federal Sector

The halo error results for the 1998 SQA Public State and Federal Sector are listed in Table 4.10.

From Table 4.7, Table 4.8, Table 4.9 and Table 4.10, halo error was present for all examiners in all four sectors, with the exception of six inconclusive cases for the private

Table 4.8. Halo Error for the 1998 SQA Private Service Sector (Sienknecht, 1999)

Examiner	Determinant of Dimension Intercorrelation Matrix	Evidence of Halo Error
1	0	yes
2	0	yes
3	0	yes
4	0	yes
5	0	yes
6	0	yes

Table 4.9. Halo Error for the 1998 SQA Public Local Sector (Sienknecht, 1999)

Examiner	Determinant of Dimension Intercorrelation Matrix	Evidence of Halo Error
1	0	yes
2	0	yes
3	0	yes
4	0	yes
5	0	yes
6	0	yes

Table 4.10. Halo Error for the 1998 SQA

Public State and Federal Sector (Sienknecht, 1999)

Examiner	Determinant of Dimension Intercorrelation Matrix	Evidence of Halo Error
1	0	yes
2	0	yes
3	0	yes
4	0	yes
5	0	yes
6	0	yes

manufacturing sector. Given the majority of determinants were effectively zeros, there were no differences in halo error between examiners.

4.1.3 Leniency or Severity for the 1998 SQA Scoring Process

Leniency or severity was addressed by computing a two-way ANOVA for each of the eight dimensions in each sector. The emergence of a statistically significant rater main effect was interpreted as evidence of leniency or severity. Leniency or severity was also assessed by a statistically significant difference between a dimension mean rating and the scale midpoint. Care should be taken in explaining the results from the latter method (see the comments in section 2.5). The leniency or severity results measured by these two methods are listed in Table 4.11 and Table 4.12.

In Table 4.11, Sienknecht calculated R-squared values. It seems strange why he calculated the R-squared values. The P values are more appropriate here.

Table 4.11. Rater Main Effect R-squared Values for all 1998 Sectors (Sienknecht, 1999)

Dimension	Rater Main Effect			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Maturity of Effort	0.762**	0.538	0.634	0.703**
2 Top Management Commitment and Involvement (Leadership)	0.687**	0.430	0.535	0.572
3 Employee Involvement, Development, and Management of Participation	0.555	0.334	0.524	0.646*
4 Recognition and Rewards Systems	0.494**	0.458	0.499	0.592
5 Plan for Continuous Improvement	0.503	0.526**	0.727**	0.473
6 Performance Measurement Process (Use of Information)	0.644**	0.576**	0.334	0.497
7 Customer and Supplier Involvement	0.551*	0.602	0.361	0.580
8 Results over Time	0.720**	0.786**	0.667**	0.644**

Note: asterisks denote level of significance (** significant at 0.05 level, * significant at 0.10 level)

Table 4.12. Mean Ratings for all 1998 Sectors (Sienknecht, 1999)

Dimension	Mean Dimension Rating			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Maturity of Effort	0.686**	0.632**	0.498	0.650**
2 Top Management Commitment and Involvement (Leadership)	0.619**	0.655**	0.567**	0.608**
3 Employee Involvement, Development, and Management of Participation	0.624**	0.638**	0.522	0.628**
4 Recognition and Rewards Systems	0.623**	0.668**	0.535	0.531
5 Plan for Continuous Improvement	0.678**	0.686**	0.519	0.622**
6 Performance Measurement Process (Use of Information)	0.661**	0.636**	0.545	0.611**
7 Customer and Supplier Involvement	0.695**	0.676**	0.517	0.580
8 Results over Time	0.647**	0.543	0.490	0.533

Note: asterisks denote level of significance (** significant at 0.05 level, * significant at 0.10 level) if mean dimension rating is statistically different from the scale midpoint (0.5), based on a two-tailed t-test.

In Table 4.12, the scale midpoint of 0.5 was assumed to be the mean of the applicant population. This may be not true for the SQA. The mean score of the SQA applicant organizations may be better than 0.5 even though no leniency exists, just because these organizations were really good.

As measured by rater main effect, leniency or severity was evident for:

- Dimensions 1, 2, 4, 6, and 8 in the private manufacturing sector;
- Dimensions 5, 6, and 8 in the private service sector;
- Dimensions 5 and 8 in the public local sector;
- Dimension 1 and 8 in the public state and federal sector.

As measured by mean dimension ratings, leniency or severity exists for the following sectors and dimensions:

- All dimensions in the private manufacturing sector;

- Dimensions 1 through 7 in the private service sector;
- Dimension 2 in the public local sector;
- Dimensions 1, 2, 3 and 5 in the public state and federal sector.

The inconsistency between the results from the rater main effect measure and those from mean dimension ratings measure is likely due to the different methods used to measure the leniency or severity. That is, the methods are not completely convergent. On an ordinal basis, the relative differences between sectors are consistent with both methods, implying a gross level of convergence.

4.1.4 Range Restriction for the 1998 SQA Scoring Process

Range restriction was addressed through the use of a variance analysis in which smaller standard deviation of ratings assigned to all ratees on a given dimension was interpreted as range restriction. The degree to which range restriction was present in the ratings was assessed by comparing the variance against the actual variance from the empirical distribution of the entire data set (0.0435 in this case) (Sienknecht, 1999).

Normally we do not know what type of distribution the score population fits, so we have no way to get the base or “true” variance. One representative of this “true” variance is the actual variance from the empirical distribution of the entire data set. The bigger the data set, the better. We could use all the available scores (all the SQA scores for different years) to get this “true” variance if feasible.

The results are listed in Table 4.13 and 4.14. So evidence of range restriction exists for the following sectors and dimensions (Sienknecht, 1999):

- All dimensions in the private manufacturing sector;
- Dimensions 1, 2, 4 and 5 in the private service sector;

Table 4.13. Variance Analysis for all 1998 Sectors (Sienknecht, 1999)

Dimension	Variance			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Maturity of Effort	3.51E-02	3.17E-02	2.84E-02	4.03E-02
2 Top Management Commitment and Involvement (Leadership)	3.67E-02	2.93E-02	2.30E-02	3.13E-02
3 Employee Involvement, Development, and Management of Participation	3.76E-02	4.98E-02	2.62E-02	2.15E-02
4 Recognition and Rewards Systems	4.27E-02	4.16E-02	4.88E-02	2.86E-02
5 Plan for Continuous Improvement	3.12E-02	4.14E-02	4.76E-02	2.89E-02
6 Performance Measurement Process (Use of Information)	3.26E-02	5.35E-02	3.79E-02	5.75E-02
7 Customer and Supplier Involvement	3.83E-02	4.52E-02	3.45E-02	6.55E-02
8 Results over Time	4.13E-02	7.10E-02	4.74E-02	6.12E-02

Note: variance value less than 4.35E-02 (based on the empirical distribution) implies range restriction.

Table 4.14. Range Restriction Results for all 1998 Sectors (Sienknecht, 1999)

Dimension	Evidence of Range Restriction			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Maturity of Effort	yes	yes	yes	yes
2 Top Management Commitment and Involvement (Leadership)	yes	yes	yes	yes
3 Employee Involvement, Development, and Management of Participation	yes	no	yes	yes
4 Recognition and Rewards Systems	yes	yes	no	yes
5 Plan for Continuous Improvement	yes	yes	no	yes
6 Performance Measurement Process (Use of Information)	yes	no	yes	no
7 Customer and Supplier Involvement	yes	no	yes	no
8 Results over Time	yes	no	no	no

- Dimensions 1, 2, 3, 6, and 7 in the public local sector;
- Dimensions 1, 2, 3, 4, and 5 in the public state and federal sector.

Sienknecht (1999) used the simple comparison of the dimension variance to the entire data variance to decide whether or not the range restriction exists. This is not a scientific way to compare two variances. Instead, the chi-square test should be used to make the comparison of the dimension variance to the entire data variance. The recalculated range restriction results for the 1998 SQA scores are listed in Table 4.15, Table 4.16, Table 4.17, Table 4.18 and Table 4.19.

Based on the recalculation, range restriction exists for the following sectors and dimensions:

- Dimensions 5 and 6 in the private manufacturing sector;
- Dimensions 2 in the public local sector;
- Dimensions 3 in the public state and federal sector.

Table 4.15. Recalculated Range Restriction Results for the 1998 Private
Manufacturing Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.0435)
1	3.51E-02	57.3052	0.1198	no
2	3.67E-02	59.8743	0.1760	no
3	3.76E-02	61.4146	0.2156	no
4	4.27E-02	69.6338	0.4764	no
5	3.12E-02	51.0722	0.0357**	yes
6	3.26E-02	53.1745	0.0564**	yes
7	3.83E-02	62.4113	0.2434	no
8	4.13E-02	67.3105	0.3978	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.16. Recalculated Range Restriction Results
for the 1998 Private Service Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.0435)
1	3.17E-02	25.5266	0.1205	no
2	2.93E-02	23.5878	0.0712*	no
3	4.98E-02	40.0303	0.7433	no
4	4.16E-02	33.4551	0.4573	no
5	4.14E-02	33.2780	0.4486	no
6	5.35E-02	43.0449	0.8352	no
7	4.52E-02	36.3660	0.5951	no
8	7.10E-02	57.1254	0.9895	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.17. Recalculated Range Restriction Results
for the 1998 Public Local Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.0435)
1	2.84E-02	14.9928	0.1051*	no
2	2.30E-02	12.1417	0.0318**	yes
3	2.62E-02	13.8398	0.0686*	no
4	4.88E-02	25.7939	0.6893	no
5	4.76E-02	25.1428	0.6570	no
6	3.79E-02	20.0256	0.3596	no
7	3.45E-02	18.2317	0.2552	no
8	4.74E-02	25.0470	0.6521	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.18. Recalculated Range Restriction Results
for the 1998 Public State and Federal Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.0435)
1	4.03E-02	15.7421	0.4578	no
2	3.13E-02	12.2087	0.2127	no
3	2.15E-02	8.4136	0.0431**	yes
4	2.86E-02	11.1618	0.1520	no
5	2.89E-02	11.2863	0.1587	no
6	5.75E-02	22.4704	0.8327	no
7	6.55E-02	25.5931	0.9178	no
8	6.12E-02	23.9004	0.8778	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.19. Recalculated Range Restriction Results for all 1998 Sectors

Dimension	Evidence of Range Restriction			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Maturity of Effort	no	no	no	no
2 Top Management Commitment and Involvement (Leadership)	no	no	yes	no
3 Employee Involvement, Development, and Management of Participation	no	no	no	yes
4 Recognition and Rewards Systems	no	no	no	no
5 Plan for Continuous Improvement	yes	no	no	no
6 Performance Measurement Process (Use of Information)	yes	no	no	no
7 Customer and Supplier Involvement	no	no	no	no
8 Results over Time	no	no	no	no

4.1.5 Overall Conclusion for the Rating Effectiveness of 1998 SQA Scoring Process

The rating effectiveness (as measured by interrater reliability, halo error, leniency or severity, and range restriction) was low to moderate. Interrater reliability was perhaps the most critical measure of rating effectiveness addressed by Sienknecht's (1999) research. Unfortunately, the state of interrater reliability for the 1998 SQA was probably less than desirable.

There was strong evidence of halo error in every measurable case. Leniency or severity was evident in the great majority of cases. Leniency or severity (measured by a comparison to scale mid-point) could be attributed not only to the examiners, but to the applicants, which were likely above average. The fact that there was strong evidence of leniency in the ratings was not necessarily problematic, assuming that the rating scale was calibrated to the whole population of potential applicants and decisions were made accordingly.

Range restriction was present in only a few cases. So range restriction was not a problem.

4.2 Interrater Reliability Results and Analysis for the 1999 SQA Scoring Process

This section addresses the three research questions for the 1999 SQA interrater reliability **Q1a: How much interrater reliability is present in the ratings? Q1b: What differences in interrater reliability exist between dimensions within a given sector? and Q1c: What differences in interrater reliability exist in a given sector between 1998 scores and 1999 scores?**

Interrater reliability is measured by the intraclass correlation coefficient ICC(1, 1) (Shrout and Fleiss, 1979) or by the standard deviations of the ratings assigned to a

Table 4.20. 1999 SQA ICC(1,1) Results for all Sectors

Dimension	ICC(1,1) and related significance level			
	Private Manufacturing Sector	Private Service Sector	Public Local Sector	Public State and Federal Sector
1 Leadership	0.359**	-0.3059	0.3746*	0.0687
2 Strategic Planning	0.2057	-0.3295	0.1439	-0.2191
3 Customer and Market Focus	0.1599	0.4502*	0.1692	0.0575
4 Information and Analysis	0.1738	-0.1667	-0.0873	-0.1956
5 Human Resource Focus	0.1453	-0.2763	-0.1818	-0.2493
6 Process Management	0.6151**	-0.2464	-0.0867	-0.121
7 Business Results	0.5047**	-0.3223	-0.3123	0.4028*

Note: significance levels (** significant at 0.05 level, * significant at 0.10 level) are based on $F = \text{BMS}/\text{WMS}$, which is used for testing the ICC(1,1) against zero.

particular rater by several raters for a given dimension (Bernardin, 1977). Boxplots are used to present the results when the standard deviation method is used.

Table 4.20 is a summary of the ICC(1, 1) results for all sectors. The computer printouts are listed in Appendix B. The following four sections present the interrater reliability results for the 1999 SQA scoring process.

4.2.1 Interrater Reliability Results for the 1999 SQA Private Manufacturing Sector

Table 4.21 summarizes the ICC(1, 1) results for the 1999 SQA Private Manufacturing Sector. Following a general practice in industrial and organizational psychology, the ICC results are classified as follows: low (less than 0.3), moderate (0.3 to 0.7), and high (0.7 to 1.0) (Sienknecht, 1999). From Table 4.21, there are differences in interrater reliability between dimensions. Dimensions 1, 6 and 7 have moderate interrater reliability (ICC values are greater than 0.3) and dimensions 2, 3, 4, and 5 have low interrater reliability (ICC values are less than 0.3 and not significantly different from zero.).

Table 4.21. ICC(1,1) for the 1999 SQA Private Manufacturing Sector

Dimension	ICC(1,1)	Significance
1 Leadership	0.359	0.0419
2 Strategic Planning	0.2057	0.1407
3 Customer and Market Focus	0.1599	0.1891
4 Information and Analysis	0.1738	0.1735
5 Human Resource Focus	0.1453	0.2067
6 Process Management	0.6151	0.0017
7 Business Results	0.5047	0.0087

Note: significance levels are based on $F = \text{BMS}/\text{WMS}$, which is used to testing the ICC(1,1) against zero.

Table 4.22. Interquartile Ranges for the 1999 SQA Private Manufacturing Sector

Ratee	Dimension							Average
	1	2	3	4	5	6	7	
1	0.37	0.5	0.3275	0.175	0.355	0.2625	0.185	0.3107
2	0.215	0.45	0.3725	0.45	0.045	0.1625	0.0775	0.2532
3	0.1575	0.35	0.295	0.2875	0.14	0.065	0.27	0.2236
4	0.32	0.325	0.43	0.25	0.135	0.49	0.475	0.3464
5	0.485	0.3	0.455	0.15	0.265	0.3	0.14	0.2993

Figure 4.1 presents the boxplots which illustrate the examiner-based variance in applicants' scores for each dimension. Table 4.22 lists the interquartile range. Small interquartile ranges are interpreted as greater interrater reliability. The cells in Table 4.22 are shaded increasingly darker with increasing interquartile range.

From Table 4.22, for the 1999 SQA Private Manufacturing Sector, 31.4% (11) of the cases have a low (0.000 to 0.199) interquartile range, 45.7% (16) of the cases have a moderate (0.200 to 0.399) interquartile range, 22.9% (8) of the cases have a high (0.400 to 0.599) interquartile range. Based on Table 4.22, approximately two thirds of the cases have moderate or high interquartile ranges and no average interquartile range for any ratee is low, so the 1999 SQA Private Manufacturing Sector has low to moderate interrater reliability.

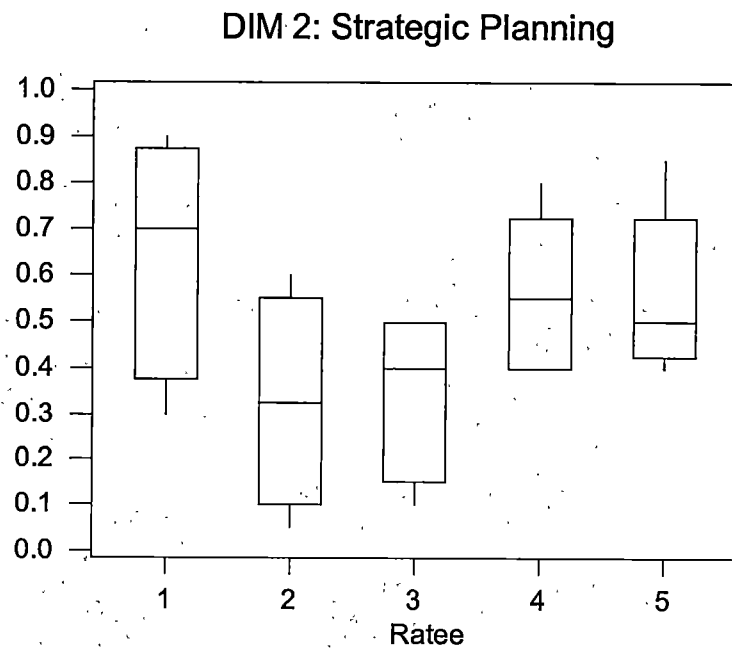
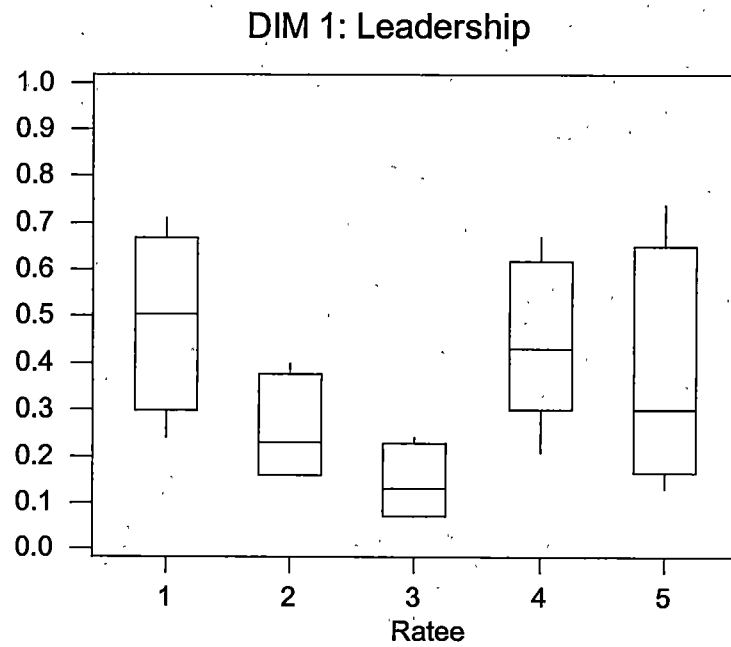
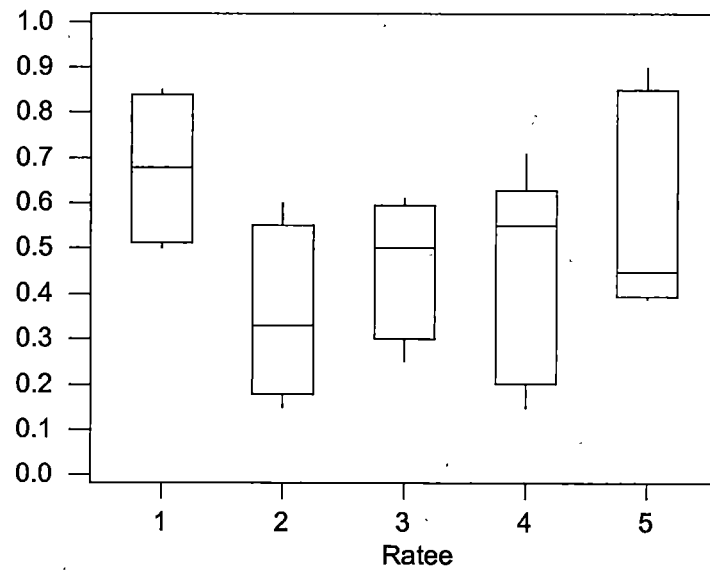


Figure 4.1. Boxplots for Interrater Reliability,
1999 SQA Private Manufacturing Sector

DIM 3: Customer and Market Focus



DIM 4: Information and Analysis

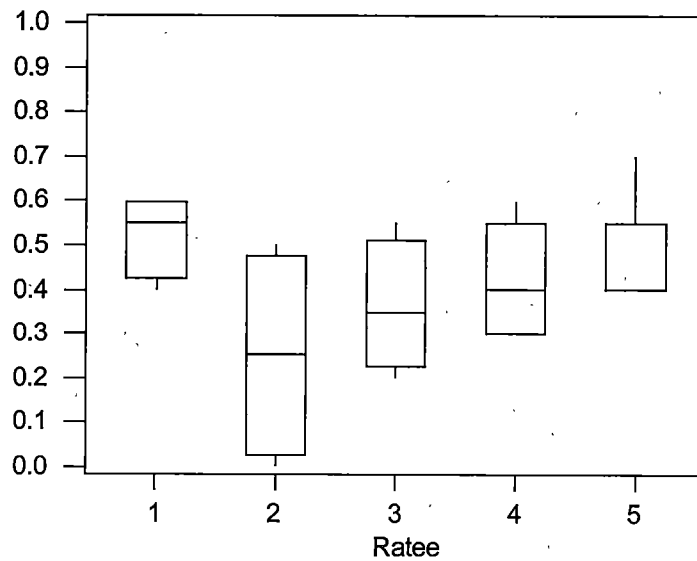
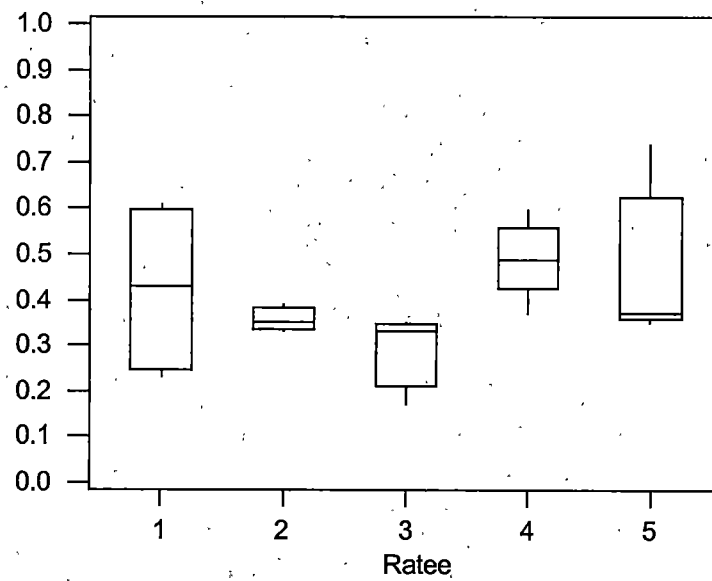


Figure 4.1. Continued

DIM 5: Human Resource Focus



DIM 6: Process Management

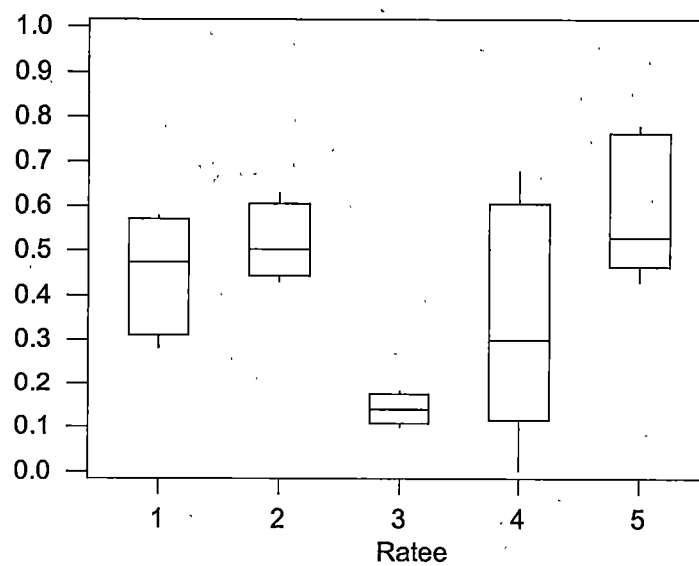


Figure 4.1. Continued

DIM 7: Business Results

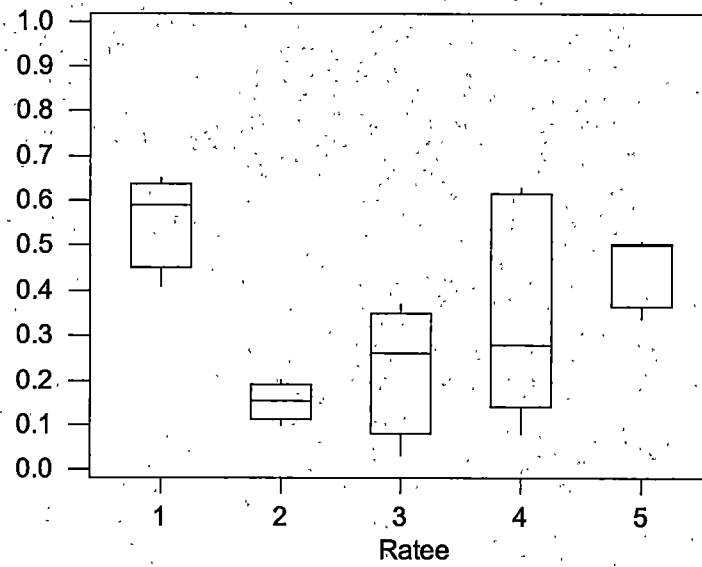


Figure 4.1. Continued

Compared with the 1998 SQA results, the Private Manufacturing Sector exhibited low to moderate interrater reliability in both 1998 and 1999 scoring processes.

4.2.2 Interrater Reliability Results for the 1999 SQA Private Service Sector

Table 4.23 summarizes the ICC(1, 1) results for the 1999 SQA Private Service Sector. From Table 4.23, only dimension 3 has moderate interrater reliability (ICC value is greater than 0.3 and significant at 0.10 level.). All the other dimensions have negative ICC values, expressing extremely low interrater reliability. So the 1999 SQA Private Service Sector has low interrater reliability.

Figure 4.2 presents the boxplots which illustrate the examiner-based variance in applicants' scores for each dimension of the 1999 SQA Private Service Sector. Table 4.24 lists the interquartile ranges for the 1999 SQA Private Service Sector.

Table 4.23. ICC(1,1) for the 1999 SQA Private Service Sector

Dimension	ICC(1,1)	Significance
1 Leadership	-0.3059	0.81
2 Strategic Planning	-0.3295	0.9294
3 Customer and Market Focus	0.4502	0.0842
4 Information and Analysis	-0.1667	0.537
5 Human Resource Focus	-0.2763	0.7269
6 Process Management	-0.2464	0.6635
7 Business Results	-0.3223	0.8793

Note: significance levels are based on $F = \text{BMS}/\text{WMS}$, which is used to testing the ICC(1,1) against zero.

Table 4.24. Interquartile Ranges for the 1999 SQA Private Service Sector

Ratee	Dimension							Average
	1	2	3	4	5	6	7	
1	0.235	0.3	0.26	0.35	0.26	0.055	0.33	0.2557
2	0.095	0.2875	0.1425	0.1	0.1675	0.1475	0.3675	0.1868

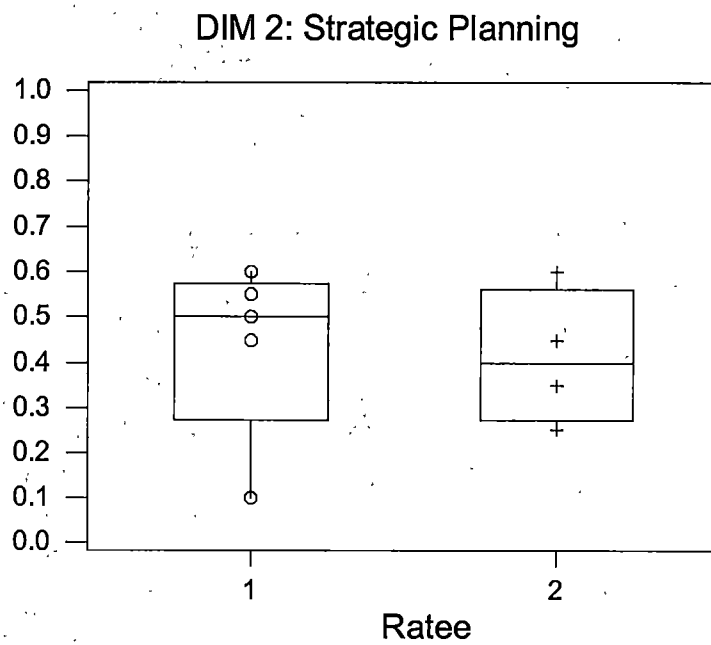
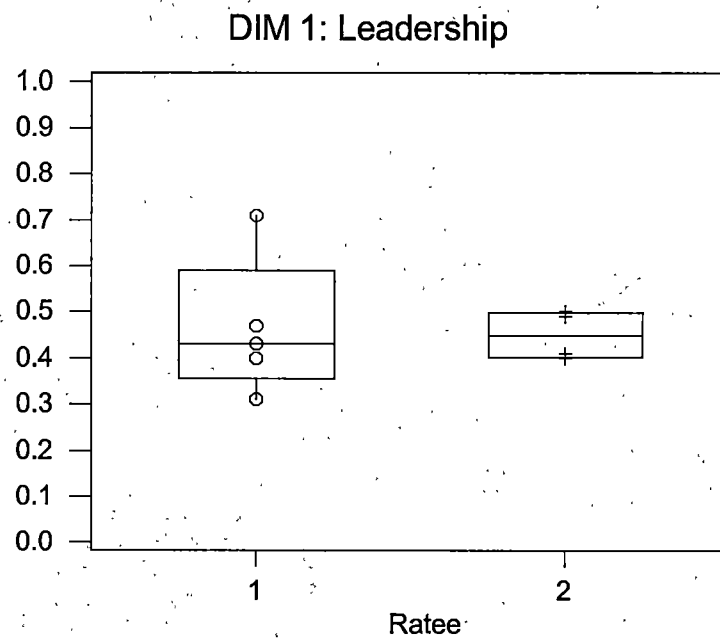
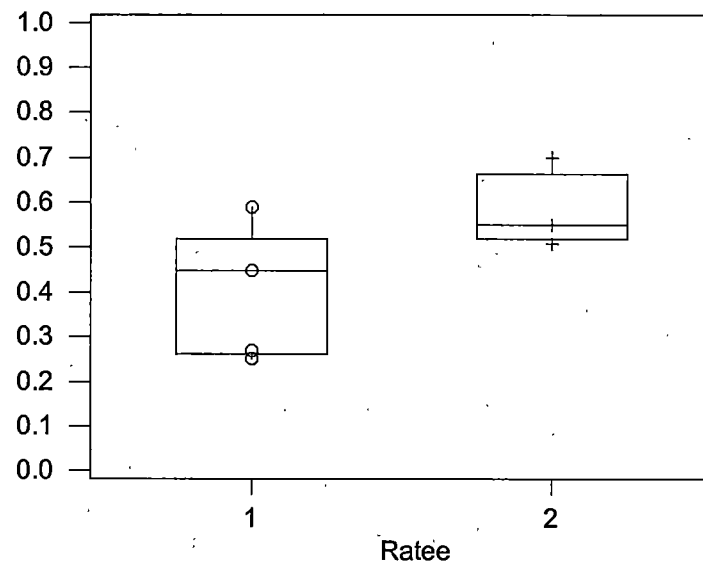


Figure 4.2. Boxplots for Interrater Reliability,
the 1999 SQA Private Service Sector

DIM 3: Customer and Market Focus



DIM 4: Information and Analysis

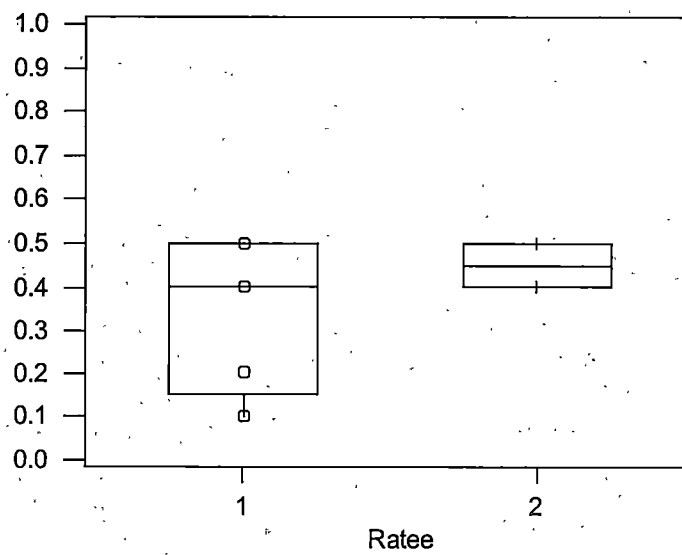
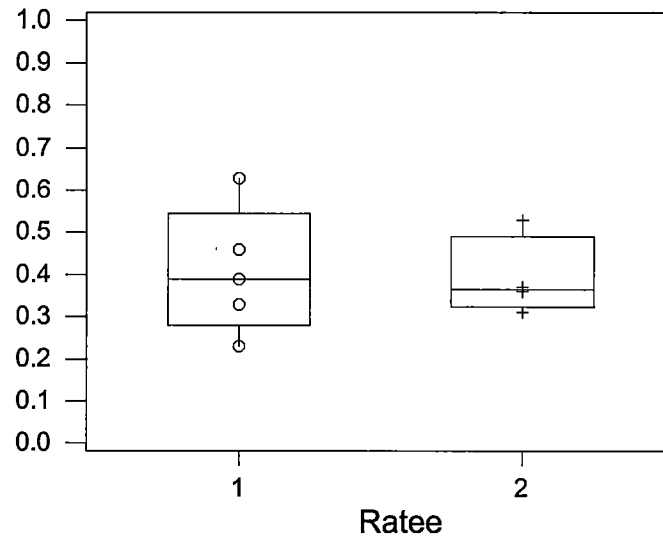


Figure 4.2. Continued

DIM 5: Human Resource Focus



DIM 6: Process Management

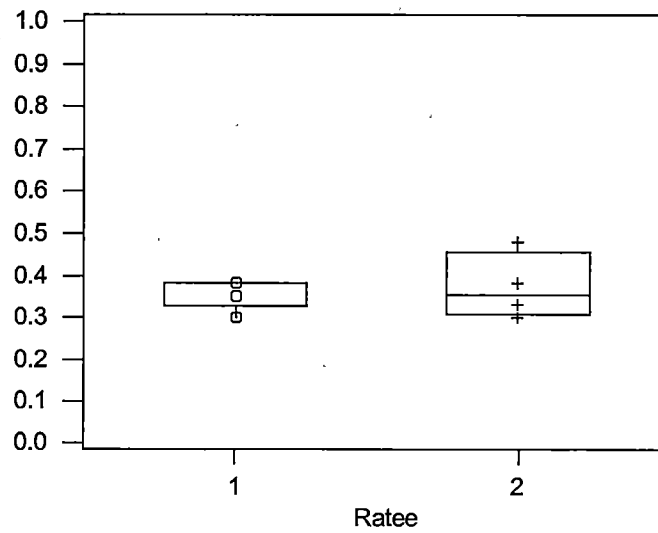


Figure 4.2. Continued

DIM 7: Business Results

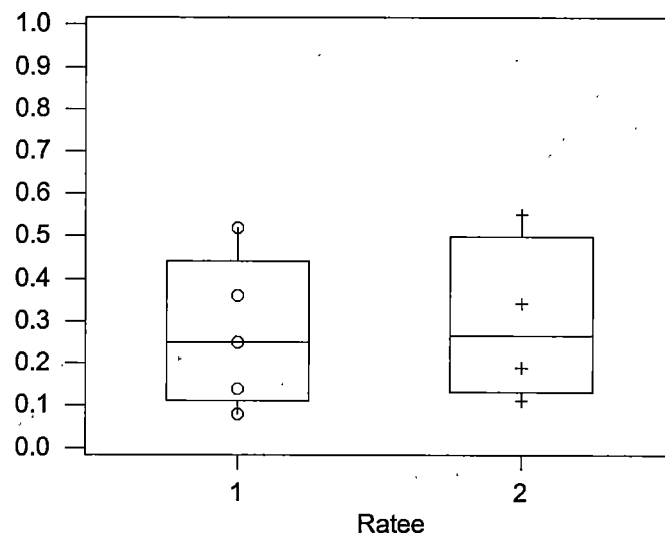


Figure 4.2. Continued

From Table 4.24, for the 1999 SQA Private Service Sector, 42.9% (6) of the cases have a low (0.000 to 0.199) interquartile range, 57.1% (8) of the cases have a moderate (0.200 to 0.399) interquartile range. Based on Table 4.24, the Private Service Sector has low to moderate interrater reliability. The inconsistency between the ICC results and the interquartile range results comes from the different methods used to measure the interrater reliability. According to the discussion in section 2.3, the intraclass correlation-generalizability theory approach seemed to offer the precision, comprehensiveness and flexibility required to deal with the complexity of reliability assessment (Berk, 1979). Based on the ICC results, the 1999 SQA Private Service Sector exhibited low interrater reliability.

Compared with the 1998 SQA results, the Private Service Sector exhibited lower interrater reliability in 1999 than in 1998. In the 1998 SQA, the same six examiners rated all six applicant organizations in the Private Service Sector, while in the 1999 SQA, the two applicant organizations were rated by two different group of examiners. For the purpose of high interrater reliability, it is better to let the same group of examiners rate all applicant organizations.

Table 4.25. ICC(1,1) for the 1999 SQA Public Local Sector

Dimension	ICC(1,1)	Significance
1 Leadership	0.3746	0.0797
2 Strategic Planning	0.1439	0.2412
3 Customer and Market Focus	0.1692	0.2177
4 Information and Analysis	-0.0873	0.5314
5 Human Resource Focus	-0.1818	0.6914
6 Process Management	-0.0867	0.5305
7 Business Results	-0.3123	0.9533

Note: significance levels are based on $F = \text{BMS}/\text{WMS}$, which is used to testing the ICC(1,1) against zero.

Table 4.26. Interquartile Ranges for the 1999 SQA Public Local Sector

Ratee	Dimension							Average
	1	2	3	4	5	6	7	
1	0.365	0.35	0.2	0.275	0.155	0.25	0.145	0.2486
2	0.185	0.125	0.13	0.2	0.2	0.11	0.22	0.1671
3	0.3275	0.35	0.275	0.3	0.3225	0.3375	0.588	0.3572

4.2.3 Interrater Reliability Results for the 1999 SQA Public Local Sector

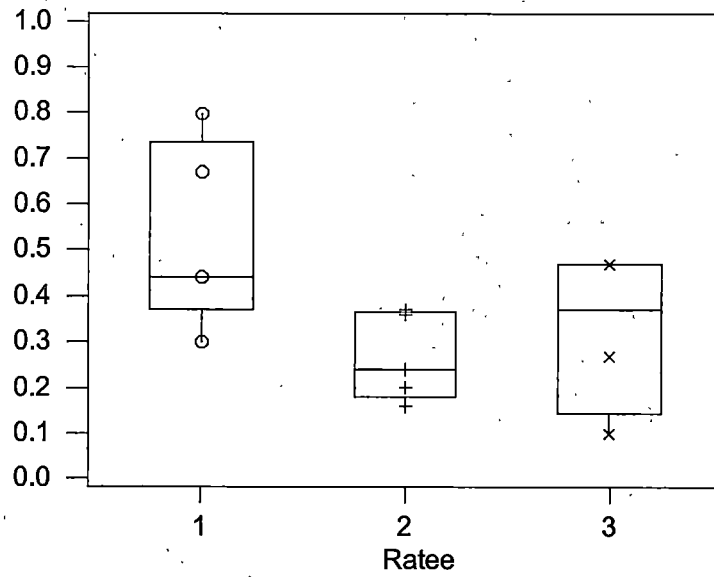
Table 4.25 summarizes the ICC(1, 1) results for the 1999 SQA Public Local Sector. From Table 4.25, only dimension 1 has moderate interrater reliability (significant at 0.10 level), and dimensions 2 and 3 have low interrater reliability. Dimensions 4, 5, 6 and 7 have negative values, expressing extremely low interrater reliability. So the individual scores of the 1999 SQA Public Local Sector exhibited low interrater reliability.

Figure 4.3 presents the boxplots which illustrate the examiner-based variance in applicants' ratings for each dimension. Table 4.26 lists the interquartile range.

From Table 4.26, for the 1999 SQA Public Local Sector, 28.5% (6) of the cases have a low (0.000 to 0.199) interquartile range, 66.7% (14) of the cases have a moderate (0.200 to 0.399) interquartile range, 4.8% (1) of the cases have a high (0.400 to 0.599) interquartile range. Based on Table 4.26, 71.5% of cases have moderate or high interquartile ranges, so the 1999 SQA Public Local Sector exhibited low interrater reliability.

Compared with the 1998 SQA results, the Public Local Sector exhibited low interrater reliability for the 1999 scoring process, while it exhibited low to moderate interrater reliability for the 1998 scoring process. The three applicant organizations in the Public Local Sector were rated by three different groups of examiners in the 1999 SQA

DIM 1: Leadership



DIM 2: Strategic Planning

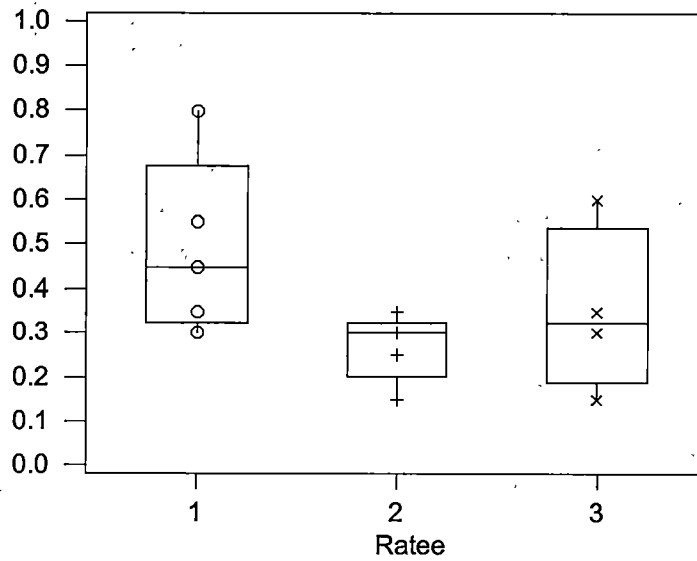
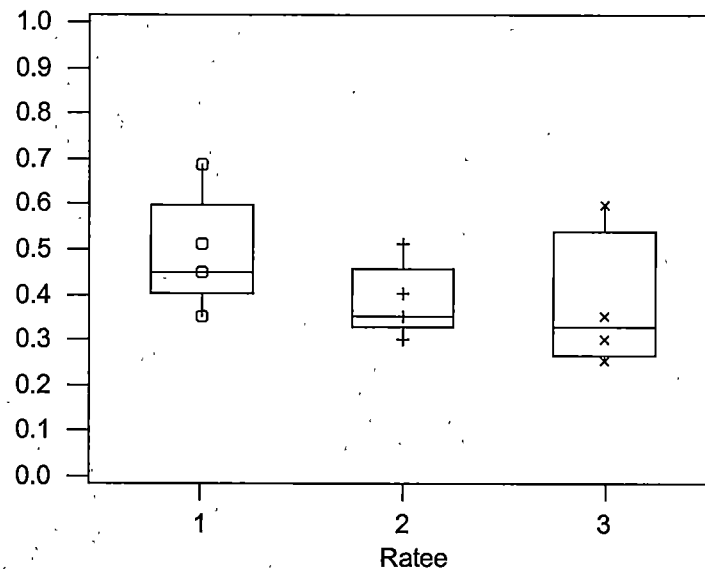


Figure 4.3. Boxplots for Interrater Reliability, the 1999 SQA Public Local Sector

DIM 3: Customer and Market Focus



DIM 4: Information and Analysis

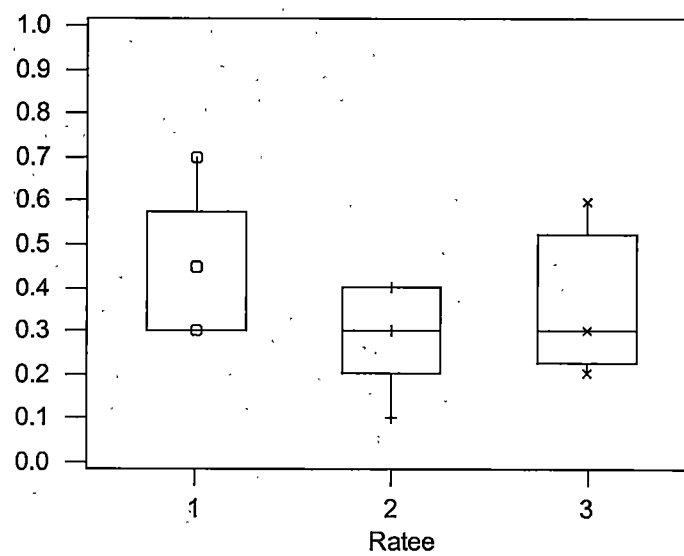
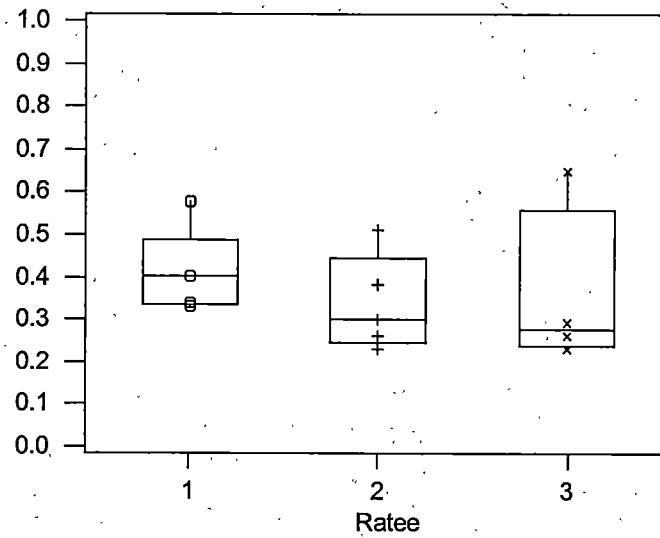


Figure 4.3. Continued

DIM 5: Human Resource Focus



DIM 6: Process Management

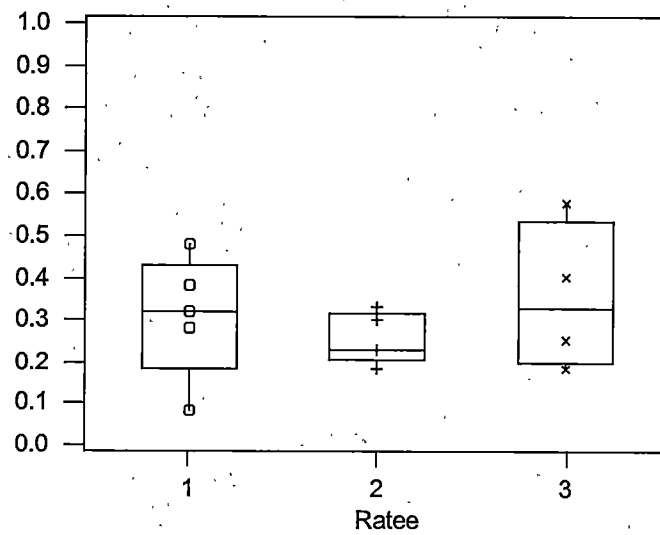


Figure 4.3. Continued

DIM 7: Business Results

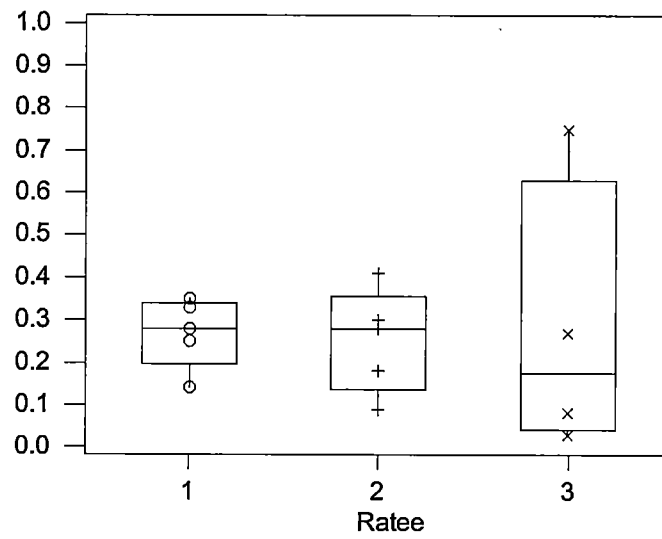


Figure 4.3. Continued

scoring process, while in the 1998 SQA scoring process, all applicants in the Public Local Sector were rated by the same group of examiners. For the purpose of high interrater reliability, it is better to have all applicants rated by the same group of examiners.

4.2.4 Interrater Reliability Results for the 1999 SQA Public State and Federal Sector

Table 4.27 summarizes the ICC(1, 1) results for the 1999 SQA Public State and Federal Sector. From Table 4.27, only dimension 7 showed moderate interrater reliability (ICC value is greater than 0.3, and significant at 0.10 level.). Dimensions 1 and 3 has low interrater reliability (ICC values are not significantly different from zero.). Dimensions 2, 4, 5, and 6 have negative ICC values, expressing extremely low interrater reliability. So the 1999 SQA Public State and Federal Sector exhibited low interrater reliability.

Figure 4.4 presents the boxplots which illustrate the examiner-based variance in applicants' ratings for each dimension. Table 4.28 lists the interquartile ranges.

Table 4.27. ICC(1,1) for the 1999 SQA Public State and Federal Sector

Dimension	ICC(1,1)	Significance
1 Leadership	0.0687	0.2757
2 Strategic Planning	-0.2191	0.7583
3 Customer and Market Focus	0.0575	0.2863
4 Information and Analysis	-0.1956	0.6808
5 Human Resource Focus	-0.2493	0.9631
6 Process Management	-0.1210	0.5166
7 Business Results	0.4028	0.0699

Note: significance levels are based on $F = \text{BMS}/\text{WMS}$, which is used to testing the ICC(1,1) against zero.

Table 4.28. Interquartile Ranges for the 1999 SQA Public State and Federal Sector

Ratee	Dimension							Average
	1	2	3	4	5	6	7	
1	0.2225	0.2125	0.35	0.375	0.375	0.44	0.3	0.325
2	0.245	0.35	0.365	0.525	0.335	0.24	0.295	0.3364

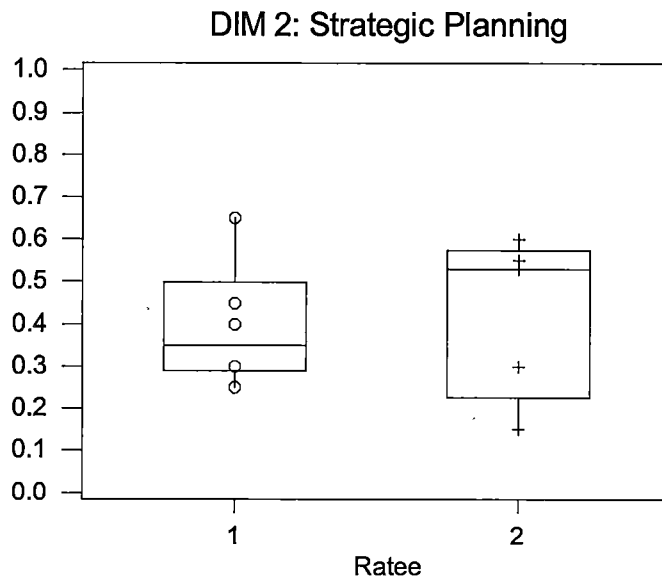
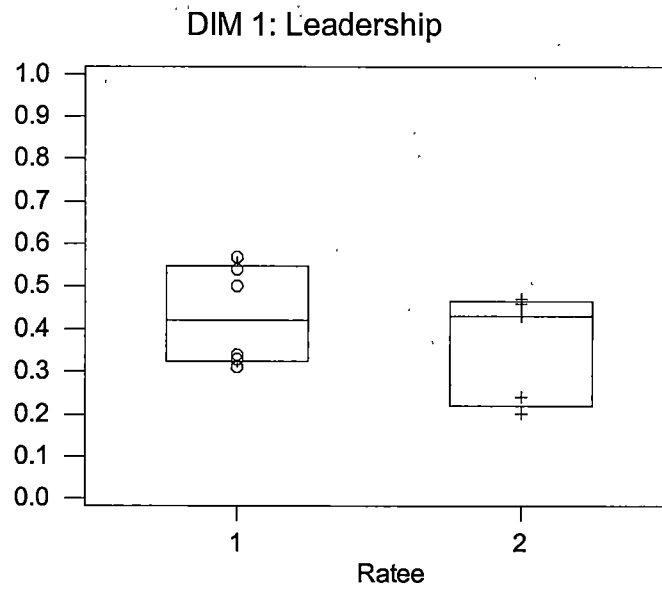
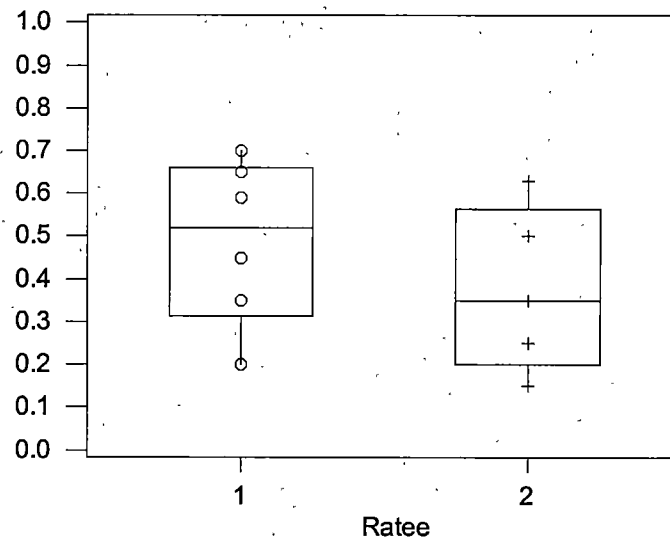


Figure 4.4. Boxplots for Interrater Reliability,
the 1999 SQA Public State and Federal Sector

DIM 3: Customer and Market Focus



DIM 4: Information and Analysis

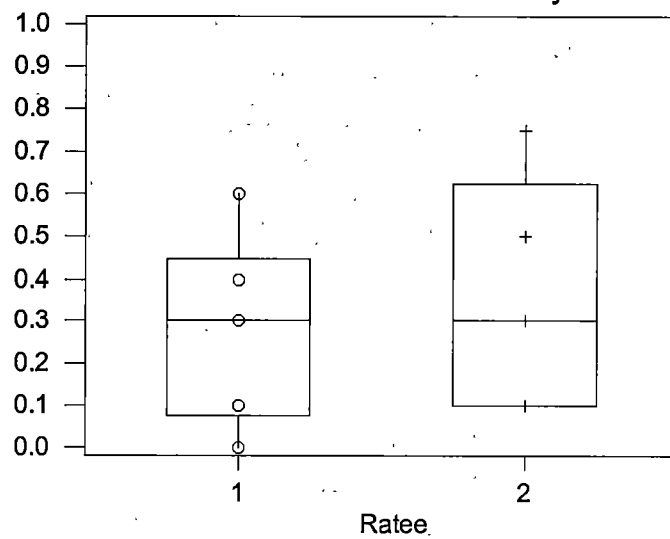
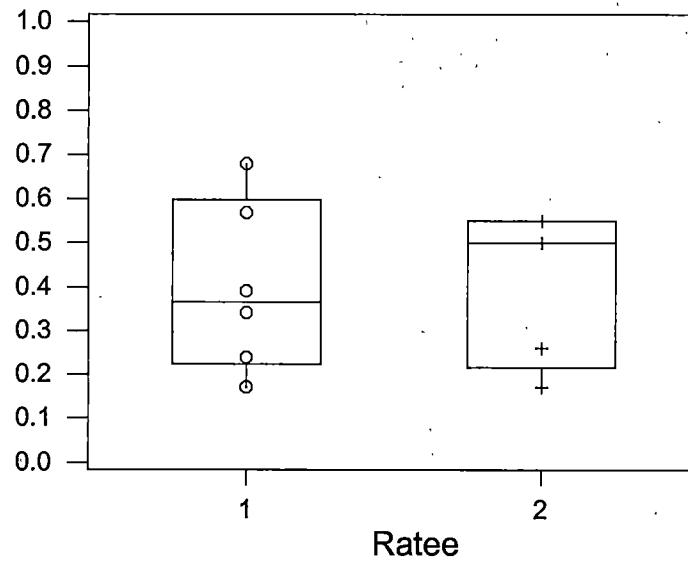


Figure 4.4. Continued

DIM 5: Human Resource Focus



DIM 6: Process Management

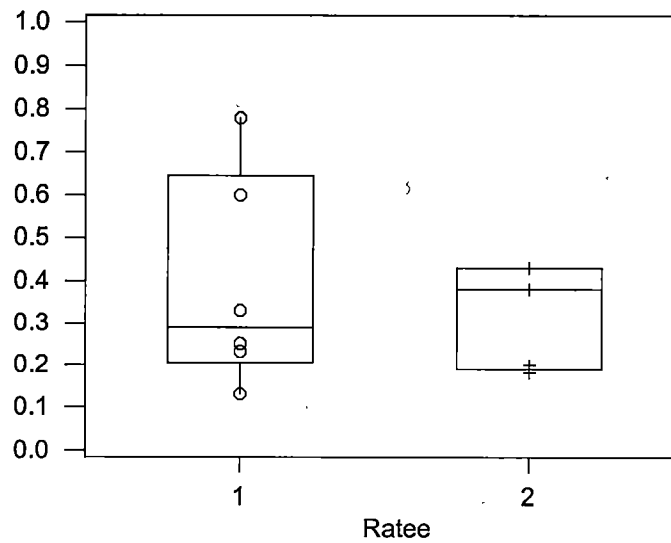


Figure 4.4. Continued

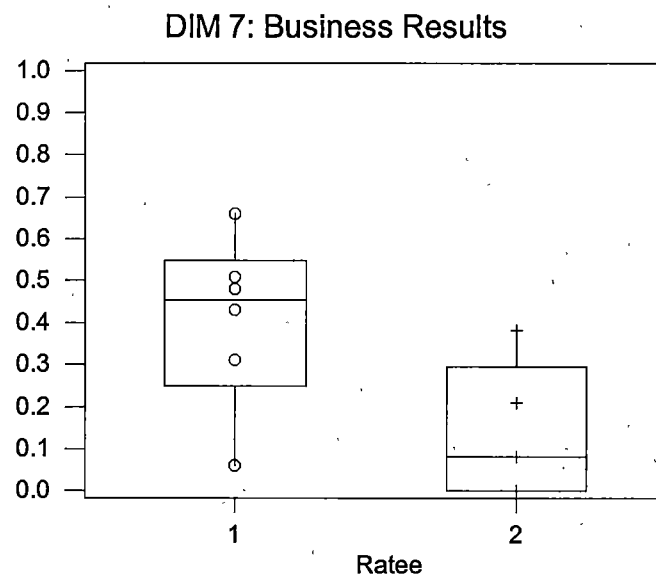


Figure 4.4. Continued

From Table 4.28, for the 1999 SQA Public State and Federal Sector, 85.7% (12) of the cases have a moderate (0.200 to 0.399) interquartile range, 14.3% (2) of the cases have a large (0.400 to 0.599) interquartile range. So the Public State and Federal Sector exhibited low interrater reliability.

Compared with the 1998 SQA results, the Public State and Federal Sector exhibited low interrater reliability for both the 1998 and 1999 scoring processes.

4.3 Halo Error Results and Analysis for the 1999 SQA Scoring Process

This section presents the results and analysis for halo error, addressing questions

Q2a: How much halo error is present in the ratings? Q2b: What differences in halo error exist between examiners within a given sector? and Q2c: What differences in halo error exist in a given sector between 1998 scores and 1999 scores?

Halo error is measured by the intercorrelations among different dimension ratings, using rater scores (over rates) for each dimension as data points (Pulakos, Schmitt and Ostroff, 1986). The determinants of the dimension intercorrelation matrices for each examiner are calculated for all the four sectors. The determinant serves as an aggregate measure of halo error for a given examiner. The determinant of a correlation matrix is small when the correlations are high. A determinant close to zero is evidence of halo error (Sienknecht, 1999). The calculation results are listed in Appendix C.

Most of the examiners rated only one or two applicants. If the examiner rated only one applicant, an intercorrelation coefficient does not exist. If the examiner rated only two applicants, the intercorrelation coefficient is always one, and the determinant will always be zero. For the 1999 SQA Private Service Sector and the Public State and Federal Sector, there are only two applicants within each of these two sectors, so the halo

Table 4.29. Halo Error for Examiners in the 1999 SQA Private Manufacturing Sector

Examiner	Determinant of Dimension Interrelation Matrix	Evidence of Halo Error
1	7.758E-14	yes
2	1.185E-16	yes

Table 4.30. Halo Error for Examiners in the 1999 SQA Public Local Sector

Examiner	Determinant of Dimension Interrelation Matrix	Evidence of Halo Error
1	does not exist	inconclusive
2	-6.856E-18	yes

error can not be measured by this method. For the 1999 SQA Private Manufacturing Sector, only two of the examiners examined more than two applicants. The halo error results for these two examiners are listed in Table 4.29. From Table 4.29, both the two examiners show halo error in their scores. For the 1999 SQA Public Local Sector, only two of the examiners rated more than two applicants, and the halo error results for these two examiners are listed in Table 4.30.

From Table 4.30, one examiner shows halo error in his scores, but it is inconclusive about the other due to the non-existence of the determinant (the examiner gave same scores to all three applicants on dimension 4, and all the correlation coefficients with this dimension could not be calculated.).

It seems the dimensional correlation method (formula (2.11)) is not so consistent with the conceptual definition of the halo error. It is suggested the variance method (formula (2.12)) be used as well. If the correlations are high and the variances are low, there is an evidence of halo error. But if both the correlations and the variances are high, there is not necessarily an evidence of halo error.

4.4 Leniency or Severity Results and Analysis for the 1999 SQA Scoring Process

This section presents the results and analysis for leniency or severity, addressing questions **Q3a: How much leniency or severity is present in the ratings? Q3b: What differences in leniency or severity exist between dimensions within a given sector?** and **Q3c: What differences in leniency or severity exist in a given sector between 1998 scores and 1999 scores?**

Leniency or severity is measured by the comparison of average dimension ratings with the midpoint of the rating scale (Bernardin, Alvares, and Cranny, 1976). Mean dimension ratings that exceed the midpoint are thought to reflect leniency, whereas mean ratings below the midpoint are thought to reflect severity. Here we are measuring the leniency or severity of the team of examiners. Leniency or severity can also be measured for an individual examiner.

4.4.1 Leniency or Severity Results and Analysis for the 1999 SQA Private Manufacturing Sector

Table 4.31 summarizes the results of the mean dimension rating data analysis. From Table 4.31, dimensions 1, 4, 5 and 7 show severity, and no dimensions show leniency. Compared with the 1998 SQA results of all dimensions showing leniency, it shows examiners in 1999 were more strict than those in 1998. This may have been caused by the SQA criteria change. The 1999 SQA criteria may have had stricter standard than the 1998 SQA criteria. The 1999 SQA criteria are in "condensed" Baldrige format. Average scores in the 30% - 40% range may be normal for the applicant pool of SQA, therefore, these scores may not suffer from a severity effect.

Table 4.31. Mean Dimension Rating Analysis for
the 1999 SQA Private Manufacturing Sector

Dimension	Descriptives			Comparison with scale midpoint (0.5)			
	Mean rating	Difference from scale midpoint	Standard deviation	t	df	Significance (2 tailed)	Evidence of leniency or severity ^a
1	0.3523	-0.1477	0.2084	-3.325	21	0.003**	yes
2	0.4955	-0.0045	0.2252	-0.095	21	0.925	no
3	0.5059	0.0059	0.2192	0.126	21	0.901	no
4	0.4068	-0.0932	0.1664	-2.627	21	0.016**	yes
5	0.4136	-0.0864	0.1358	-2.984	21	0.007**	yes
6	0.4168	-0.0832	0.2185	-1.785	21	0.089*	no
7	0.3545	-0.1455	0.1957	-3.487	21	0.002**	yes

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 2-tailed t-test significance @ 0.05.

Table 4.32. Mean Dimension Rating Analysis for
the 1999 SQA Private Service Sector

Dimension	Descriptives			Comparison with scale midpoint (0.5)			
	Mean rating	Difference from scale midpoint	Standard deviation	t	df	Significance (2 tailed)	Evidence of leniency or severity ^a
1	0.4578	-0.0422	0.1108	-1.144	8	0.286	no
2	0.4278	-0.0722	0.1679	-1.29	8	0.233	no
3	0.48	-0.02	0.1458	-0.412	8	0.691	no
4	0.3889	-0.1111	0.1453	-2.294	8	0.051**	yes
5	0.4011	-0.0989	0.1214	-2.444	8	0.04**	yes
6	0.3644	-0.1356	0.0548	-7.421	8	0.000**	yes
7	0.2822	-0.2178	0.1725	-3.788	8	0.005**	yes

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 2-tailed t-test significance @ 0.05.

4.4.2 Leniency or Severity Results and Analysis for the 1999 SQA Private Service Sector

Table 4.32 summarizes the results of the mean dimension rating data analysis. From Table 4.32, dimensions 4, 5, 6 and 7 show severity, and no dimensions show leniency. Compared with the 1998 SQA results (all except one dimension showing leniency), it shows examiners in 1999 were more strict than those in 1998. This may be also caused by the SQA criteria change.

4.4.3 Leniency or Severity Results and Analysis for the 1999 SQA Public Local Sector

Table 4.33 summarizes the results of the mean dimension rating data analysis. From Table 4.33, all dimensions show severity. Compared with the 1998 SQA results of no dimensions showing severity, it shows examiners in 1999 were more strict than those in 1998. This may have also been caused by the SQA criteria change.

4.4.4 Leniency or Severity Results and Analysis for the 1999 SQA Public State and Federal Sector

Table 4.34 summarizes the results of the mean dimension rating data analysis. From Table 4.34, dimensions 1, 4, 6 and 7 show severity, and no dimensions show leniency.

Table 4.33. Mean Dimension Rating Analysis for the 1999 SQA Public Local Sector

Dimension	Descriptives			Comparison with scale midpoint (0.5)			
	Mean rating	Difference from scale midpoint	Standard deviation	t	df	Significance (2 tailed)	Evidence of leniency or severity ^a
1	0.3779	-0.1221	0.1925	-2.374	13	0.034**	yes
2	0.3714	-0.1286	0.1773	-2.714	13	0.018**	yes
3	0.4186	-0.0814	0.1243	-2.45	13	0.029**	yes
4	0.3536	-0.1464	0.1525	-3.593	13	0.003**	yes
5	0.3686	-0.1314	0.1304	-3.77	13	0.002**	yes
6	0.3014	-0.1986	0.1296	-5.733	13	0.000**	yes
7	0.2671	-0.2329	0.1781	-4.891	13	0.000**	yes

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 2-tailed t-test significance @ 0.05.

Table 4.34. Mean Dimension Rating Analysis for
the 1999 SQA Public State and Federal Sector

Dimension	Descriptives			Comparison with scale midpoint (0.5)			
	Mean rating	Difference from scale midpoint	Standard deviation	t	df	Significance (2 tailed)	Evidence of leniency or severity ^a
1	0.3991	-0.1009	0.1225	-2.732	10	0.021**	yes
2	0.4073	-0.0927	0.1609	-1.912	10	0.085*	no
3	0.4382	-0.0618	0.1918	-1.069	10	0.31	no
4	0.3136	-0.1864	0.2346	-2.634	10	0.025**	yes
5	0.4018	-0.0982	0.178	-1.829	10	0.097*	no
6	0.3582	-0.1418	0.1962	-2.397	10	0.037**	yes
7	0.2836	-0.2164	0.2282	-3.145	10	0.010**	yes

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 2-tailed t-test significance @ 0.05.

Compared with the 1998 SQA results (5 dimensions show leniency and no dimensions show severity), it shows the examiners in 1999 were more strict than those in 1998. This may be also caused by the SQA criteria change. The 1999 SQA criteria may have more strict requirements than the 1998 SQA criteria.

4.5 Range Restriction Results and Analysis for the 1999 SQA Scoring Process

This section presents the results and analysis for range restriction, addressing questions **Q4a: How much range restriction is present in the ratings?** **Q4b: What differences in range restriction exist between dimensions within a given sector?** and **Q4c: What differences in range restriction exist in a given sector between 1998 scores and 1999 scores?**

Range restriction is measured by the variance of the ratings assigned to all applicants on a particular performance dimension (Borman and Dunnette, 1975). Smaller variance reflects greater range restriction. The variance of the entire data set is used as the standard, and if the variance of a particular performance dimension is less than this

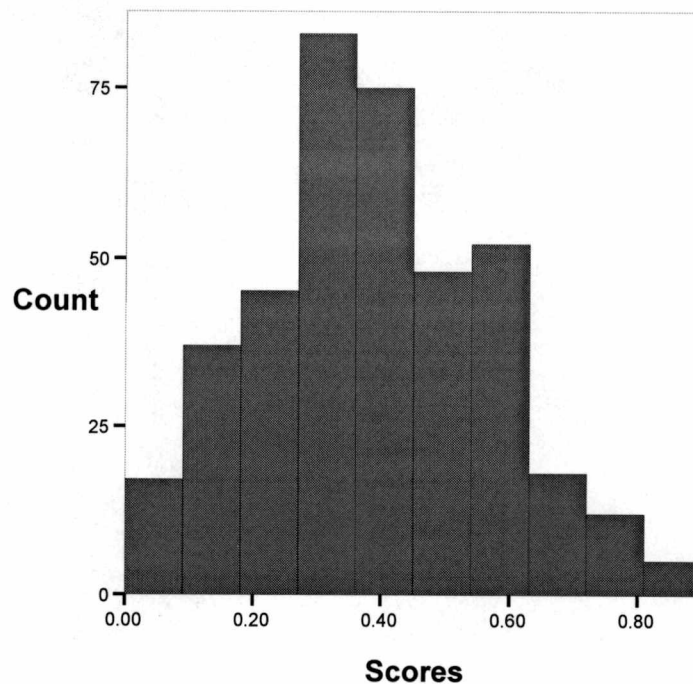


Figure 4.5. Histogram of Entire Data Set

standard, then the range restriction exists for that dimension. Figure 4.5 is the histogram of the entire data set, and the variance of the entire data set is 0.034.

4.5.1 Range Restriction Results and Analysis for the 1999 Private Manufacturing Sector

Table 4.35 summarizes the results of the range restriction for the 1999 SQA Private Manufacturing Sector. From Table 4.35, only dimension 5 shows range restriction. Compared with the 1998 SQA results of only two dimensions showing range restriction, there is no appreciable difference in range restriction for the 1998 and 1999 scoring processes.

Table 4.35. Range Restriction Results for the 1999 SQA

Private Manufacturing Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.034)
1	0.0434	27.2861	0.8324	no
2	0.0507	31.8575	0.9395	no
3	0.0480	30.1872	0.9117	no
4	0.0277	17.4012	0.3135	no
5	0.0184	11.5846	0.05**	yes
6	0.0478	30.0120	0.9083	no
7	0.0383	24.0588	0.7098	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.36. Range Restriction Results for the 1999 SQA Private Service Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.034)
1	0.0123	2.9374	0.0618*	no
2	0.0282	6.75	0.4362	no
3	0.0213	5.0874	0.2518	no
4	0.0211	5.0542	0.2482	no
5	0.0147	3.5279	0.103	no
6	0.0030	0.7189	0.0005**	yes
7	0.0297	7.121	0.4764	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

Table 4.37. Range Restriction Results for the Public Local Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.034)
1	0.0370	14.4134	0.6546	no
2	0.0314	12.2269	0.4909	no
3	0.0155	6.0143	0.0544**	yes
4	0.0233	9.0472	0.2306	no
5	0.0170	6.6188	0.0793*	no
6	0.0168	6.5350	0.0755*	no
7	0.0317	12.344	0.5003	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

4.5.2 Range Restriction Results and Analysis for the 1999 SQA Private Service Sector

Table 4.36 summarizes the results of the range restriction for the 1999 SQA Private Service Sector. From Table 4.36, only dimension 6 shows range restriction. Compared with the 1998 SQA results of no dimensions showing range restriction, there is no appreciable difference in range restriction for the 1998 and 1999 scoring processes.

4.5.3 Range Restriction Results and Analysis for the 1999 SQA Public Local Sector

Table 4.37 summarizes the results of the range restriction for the 1999 SQA Public Local Sector. From Table 4.37, only dimension 3 shows range restriction. Compared with the 1998 SQA results of only one dimension showing range restriction, there is no appreciable difference in range restriction for the 1998 and 1999 scoring processes.

4.5.4 Range Restriction Results and Analysis for the 1999 SQA Public State and Federal Sector

Table 4.38 summarizes the results of the range restriction for the 1999 SQA Public State and Federal Sector. From Table 4.38, no dimensions shows range restriction. Compared with the 1998 SQA results of only one dimension showing range restriction, there is no appreciable difference in range restriction for the 1998 and 1999 scoring processes.

Table 4.38. Range Restriction Results for
the 1999 SQA Public State and Federal Sector

Dimension	Variance	Chi-Square	Significance	Evidence of Range Restriction ^a (dimension variance < 0.034)
1	0.0150	4.4916	0.0775*	no
2	0.0259	7.7454	0.3463	no
3	0.0368	11.0117	0.6434	no
4	0.0550	16.4729	0.9131	no
5	0.0317	9.4854	0.5133	no
6	0.0385	11.5204	0.6816	no
7	0.0521	15.5811	0.8877	no

**significant at 0.05 level, *significant at 0.10 level; ^aBased on 1-tailed chi-square test significance @ 0.05.

CHAPTER V

CONCLUSIONS

Rating effectiveness for the State Quality Award scoring process was measured by interrater reliability, halo error, leniency or severity, and range restriction. Based on these four measures, the overall rating effectiveness for the 1999 SQA scoring process was low to moderate.

Interrater reliability was low to moderate for the Private Manufacturing Sector in both 1998 and 1999 SQA scoring processes. The Private Service Sector had low to moderate interrater reliability in the 1998 SQA scoring process, but had low interrater reliability in the 1999 SQA scoring process. The Public Local Sector had low to moderate interrater reliability for the 1998 SQA scoring process, but had low interrater reliability in the 1999 SQA scoring processes. The Public State and Federal Sector had low interrater reliability for both the 1998 and 1999 SQA scoring processes. The differences in the interrater reliability between 1998 and 1999 SQA scoring processes for the Private Service Sector and the Public Local Sector may have been caused by the different composition of the examiner teams. The applicant organizations in these two sectors were rated by different teams of examiners in the 1999 SQA scoring process, while in the 1998 SQA scoring process, all applicant organizations were rated by the same team of examiners. For the purpose of high interrater reliability, it is better to have all applicants rated by the same team of examiners.

There were differences in interrater reliability between dimensions, but no pattern was evident. No dimensions appeared to exhibit consistent low or high interrater reliability.

During the 1999 SQA scoring process, most of the examiners rated only one or two applicants, and halo error could not be measured for these examiners. For the 1999 SQA scoring process, halo error existed for all the examiners whose halo error could be measured. Considering all examiners in all the four sectors showing halo error in the 1998 SQA scoring process, halo error was present for all examiners and there were no differences in halo error between examiners.

In this research and Sienknecht's (1999) research, the determinant of the dimension correlation matrix for each examiner serves as an aggregate measure of halo error. It seems there is a problem with using the determinant close to zero as an evidence of halo error. For example, the determinant of the dimension correlation matrix for examiner 1 of the 1999 SQA Private Manufacturing Sector was 7.758 E-14 , which implied halo error for this examiner. Figure 5.1 shows the scores of this examiner. Based on Figure 5.1, it seems there is little evidence of obvious halo error for this examiner. That is, the examiner appears to distinguish among levels of performance exhibited on different performance dimensions. Figure 5.1 exhibits evidence of correlation between dimensions, and this is consistent with the current measure of halo error (formula 2.11). It seems the dimensional correlation method (formula 2.11) is not consistent with the conceptual definition of halo error. It is suggested that a more discriminating way to explain the correlation results for halo error is needed in future research.

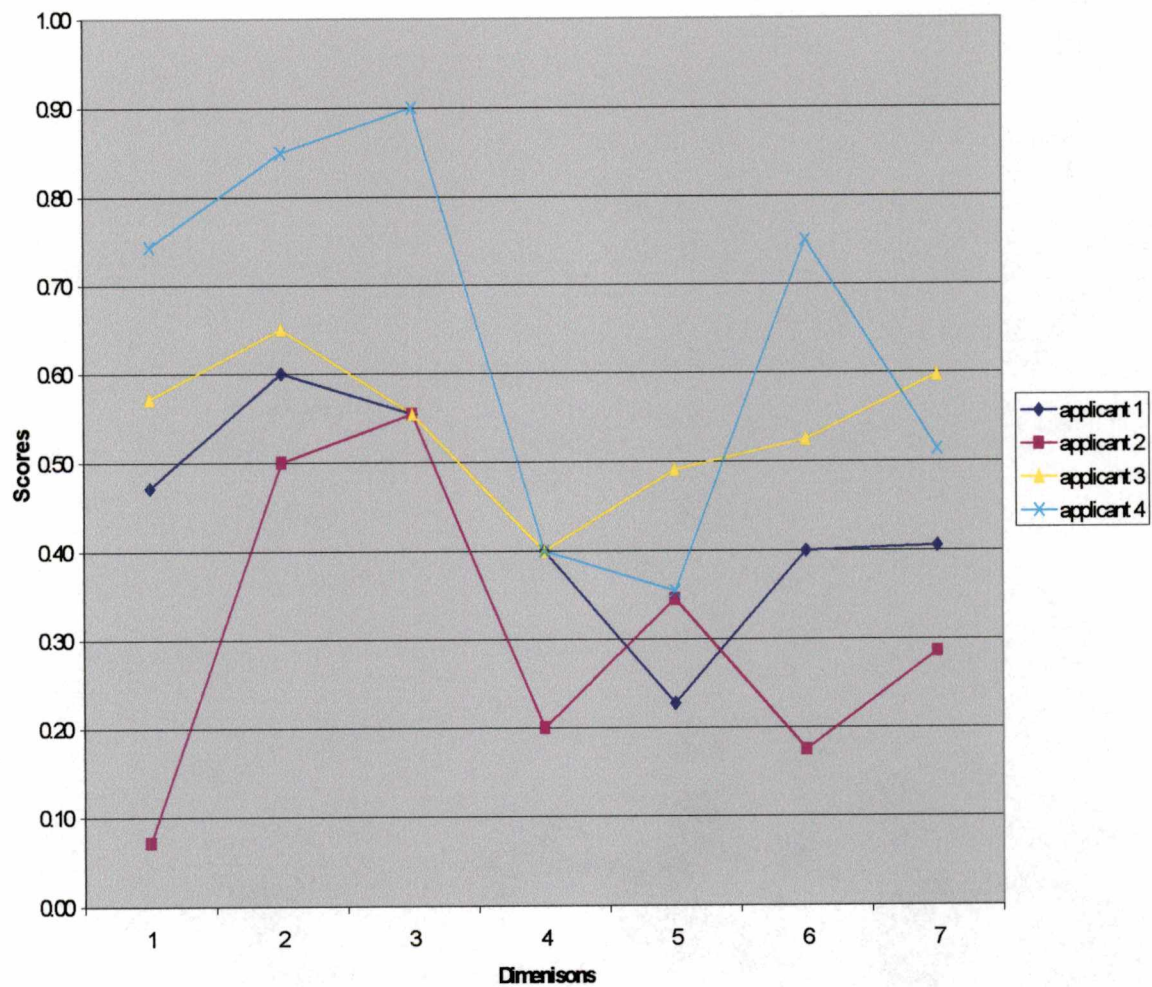


Figure 5.1. Examiner 1's Scores

For the 1999 SQA scoring process, there was no evidence of leniency by the examiners in any of the four sectors. Compared with the 1998's evidence of leniency, the examiners in the 1999 SQA scoring process assigned scores more severely than those in the 1998 SQA scoring process. For the Private Manufacturing Sector, the team of examiners exhibited severity for four dimensions in the 1999 SQA scoring process, while the team of examiners exhibited leniency for all dimensions in the 1998 SQA scoring process. For the Private Service Sector, the team of examiners exhibited severity for four dimensions in the 1999 SQA scoring process, while the team of examiners exhibited leniency for seven dimensions in the 1998 SQA scoring process. For the Public Local Sector, the team of examiners exhibited severity for all dimensions in the 1999 SQA scoring process, while the team of examiners exhibited no severity and exhibited leniency for one dimension in the 1998 SQA scoring process. For the Public State and Federal Sector, the team of examiners exhibited severity for four dimensions in the 1999 SQA scoring process, while the team of examiners exhibited leniency for five dimensions in the 1998 SQA scoring process.

The differences in the severity or leniency results between 1998 and 1999 SQA scoring processes may have been caused by the SQA criteria change from 1998 to 1999. The 1999 SQA criteria were in a "condensed Baldrige" format. These new criteria may imply a higher standard than the 1998 SQA criteria. The more information which the examiners can get from the "condensed Baldrige" format criteria may lead the examiners to give scores more severely. Also the difference of scoring guidelines between 1998 and 1999 SQA criteria may lead the examiners to give scores less leniently in the 1999 SQA scoring process. The training which examiners received may also have some effect on the

differences. In 1998, examiners received 4 hours of training, and approximately 75% of the examiners attended the training. In 1999, examiners received 8 hours of training, and approximately 75% of the examiners attended the training. The more training the examiners got, the less lenient they would be.

Scores for most dimensions in all the four sectors did not show range restriction in either the 1998 or 1999 SQA scoring processes. For the Private Manufacturing Sector, the team of examiners exhibited range restriction for only one dimension in the 1999 SQA scoring process, while the team of examiners exhibited range restriction for two dimensions in the 1998 SQA scoring process. For the Private Service Sector, the team of examiners exhibited range restriction for only one dimension in the 1999 SQA scoring process, while the team of examiners exhibited no range restriction for any dimensions in the 1998 SQA scoring process. For the Public Local Sector, the team of examiners exhibited range restriction for only one dimension in both the 1998 and 1999 SQA scoring processes. For the Public State and Federal Sector, the team of examiners exhibited no range restriction for any dimensions in the 1999 SQA scoring process, while the team of examiners exhibited range restriction for only one dimension in the 1998 SQA scoring process. No pattern was evident regarding which dimensions' scores exhibited range restriction.

Leniency or severity and range restriction were measured for the team of examiners for each criteria dimension by sector. They can also be measured for each examiner. It is suggested that in a future study, the leniency or severity and range restriction be measured for both the team of examiners and a single examiner.

This rating effectiveness analysis and evaluation was based on two consecutive years' data. The overall SQA rating effectiveness was low to moderate based on the measures of interrater reliability, halo error, leniency or severity, and range restriction. This fact gives the SQA opportunities for future improvement.

Due to the small data base and lack of control of the factors (raters, ratees, award criteria, and rating process), it is difficult to make general conclusions about the effects of these factors on rating effectiveness or to extend the findings from this research to other quality awards. But some suggestions can still be made:

- (1) For the purpose of high interrater reliability, it is better to have all applicants in a sector rated by the same team of examiners if possible. When this is not possible (e.g., the number of applicants is too large), the applicants can be divided into groups, and each group can be rated by the same team of examiners.
- (2) For the halo error, the correlation method and the variance method can be used simultaneously. If the correlations are high and the variances are low, there is an evidence of halo error. But if both the correlations and the variances are high, there is not necessarily an evidence of halo error.
- (3) Leniency or severity can be measured for both the team of examiners and a single examiner. It seems examiners give scores more severely if they get more training and are more familiar with the criteria.
- (4) Range restriction can be measured for both the team of examiners and a single examiner. It seems range restriction is not a problem for the current scoring process.
- (5) Given the small data base of this study, caution should be taken in explaining and generalizing from the statistics calculation as measure of rating effectiveness. Some

simple methods, such as charts, figures, simple comparison, etc. should also be used in explaining and generalizing the results.

- (6) Future research should be performed on a larger data base (including scores from Malcolm Baldrige National Quality Award and other state quality awards if possible), so the findings can be used more generally.

REFERENCES

REFERENCES

- Bartlett, C.J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology*, 68, 218-226.
- Bemowski, K. and Stratton, B. (1995). How do people use the Baldrige Award criteria? *Quality Progress*, 28(5), 43-47.
- Berk, R.A. (1979). Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency*, Vol. 83, No. 5, 460-472.
- Bernardin, H.J. (1977). Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology*, Vol. 62, No. 4, 422-427.
- Bernardin, H.J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, Vol. 63, No. 3, 301-308.
- Bernardin, H.J., Alvares, K.M., and Cranny, C.J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, 61, 564-570.
- Bernardin, H.J. and Pence, E.C. (1980). Effects of rater training: create new response sets and decreasing accuracy. *Journal of Applied Psychology*, Vol. 65, No. 1, 60-66.
- Bernardin, H.J., and Walter, C.S. (1977). Effects of training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*, Vol. 62, No. 1, 64-69.
- Blanz, F., and Ghiselli, E.E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology*, 25, 185-199.

- Borman, W.C. (1975). Effects of instructions to avoid halo effect on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, Vol. 60, No. 5, 556-560.
- Borman, W.C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, Vol. 64, No. 4, 410-421.
- Borman, W.C., and Dunnette, M.D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, 60, 561-565.
- Cascio, W.F., and Valenzi, E.R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology*, 62, 278-282.
- Coleman, G.D. (1996). Estimating the impact of third-party evaluator training and characteristics on the scoring of written organizational self-assessments (Doctoral dissertation, Virginia Polytechnic Institute and State University, 1996). *Dissertation Abstracts International*, 57-07B.
- Coleman, G.D., Koelling, C. P. (1998). Estimating the consistency of third-party evaluator scoring of organizational self-assessments. *Quality Management Journal*, Vol. 5, No. 3, 31-53.
- Conti, Tito (1997), Organizational Self-assessment, *Chapman & Hall*.
- Das, B., Smith, D.R., and Yeager, R.J. (1991). Personal factors affecting performance rating ability. *IIE transactions*, Vol. 23, No. 3, 267-272.
- DeCotiis, T.A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, 19, 247-266.

- Dickinson, T.L., and Tice, T.E. (1977). The discriminant validity of scales developed by retranslation. *Personnel Psychology*, 30, 255-268.
- Friedman, B.A. and Cornelius, E.T., III. (1976). Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. *Journal of Applied Psychology*, 61, 210-216.
- Futrell, D. (1995). When quality is a matter of taste, use reliability indexes. *Quality Press*, Vol. 28, No. 5, 81-86.
- Hahn, D.C., and Dipboye, R.L. (1988). Effects of training and information on the accuracy and reliability of job evaluations. *Journal of Applied Psychology*, Vol. 73, No. 2, 146-153.
- Hakes, Chris (1996). The corporate self assessment handbook, for measuring business excellence. *Chapman & Hall*.
- Heneman, H.G., III, Schwab, D.P., Huett, D.L., and Ford, J.J. (1975). Interviewer validity as a function of interview structure, biographical data, and interview order. *Journal of Applied Psychology*, 60, 748-753.
- Ivancevich, J.M. (1984). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, Vol. 69, No. 1, 85-98.
- Keinath, B.J., Gorski, B.A. (1999). An empirical study of the Minnesota Quality Award evaluation process. *Quality Management Journal*, Vol. 6, No. 1, 29-38.
- Kraut, A.I. (1975). Prediction of managerial success by peer and training-staff ratings. *Journal of Applied Psychology*, 60, 14-19.

- Kozlowski, Steve W.J. Kirsch, M.P., and Chao, G.T. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: an exploration. *Journal of Applied Psychology*, Vol. 71, No. 1, 45-49.
- Landy F.J., and Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- Landy F.J., Farr, J.L., Saal, F.E., and Freytag, W.R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*. 61, 750-758.
- Latham, G.P., and Wexley, K.N. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, Vol. 60, No. 5, 550-555.
- Murphy, K.R., and Balzer, W.K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: consequences for rating accuracy. *Journal of Applied Psychology*, Vol. 71, No. 1, 39-44.
- Murphy, K.R., Balzer, W.K., Lockhart, M.C. and Eisenman, E.J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology*, Vol. 70, No. 1, 72-84.
- Murphy, K.R. and Reynolds, D.H. (1988). Does ture halo affect observed halo? *Journal of Applied Psychology*, Vol. 73, No. 2, 235-238.
- National Institute of Standards and Technology (1998). Malcolm Baldrige National Quality Award: Why apply?
- National Institute of Standards and Technology (1997). Malcolm Baldrige National Quality Award: 1997 handbook for the board of examiners.

- Nisbett, R.E. and Wilson, T.D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social Psychology*, Vol. 35, No. 4, 250-256.
- Pulakos, E.D. (1984). A comparison of rater training programs: error training and accuracy training. *Journal of Applied Psychology*, Vol. 69, No. 4, 581-588.
- Pulakos, E.D., Schmitt, N. and Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology*, Vol. 71, No. 1, 29-32.
- Rothstein, H.R. (1990). Interrater reliability of job performance ratings: growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, Vol. 75, No. 3, 322-327.
- Saal, F.E., Downey, R.G. and Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, Vol. 88, No. 2, 413-428.
- Saal, F.E., and Landy, F.J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance*, 18, 19-35.
- Shrout, P.E., and Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, Vol. 86, No. 2, 420-428.
- Sienknecht, R.T. (1999). An empirical analysis of rating effectiveness for a state quality award. *Unpublished master's thesis, Virginia Polytechnic Institute and State University, Virginia.*
- State Quality Award (1999). State Quality Award Criteria.
- Tennessee Quality Award (1998). Handbook for the board of examiners.

- Towstropiat, O. (1984). A review of reliability procedures for measuring observer agreement. *Contemporary Educational Psychology*, 9, 333-352.
- Van der Wiele, T., Williams, R., Kolb, F. and Dale, B. (1995). Assessor training for the European Quality Award: an examination. *QW TS, March, 1995*.
- Warmke, D.L. and Billings, R.S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. *Journal of Applied Psychology*, Vol. 64, No. 2, 124-131.
- Zedeck, S., and Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, Vol. 67, No. 6, 752-758.

APPENDICES

APPENDIX A 1999 SQA RAW SCORES

This appendix contains the raw scores for the 1999 SQA scoring process.

Raw Scores for the 1999 SQA Private Manufacturing Sector

Ratee	Rater	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7
1	1	0.47	0.60	0.55	0.40	0.23	0.40	0.41
1	9	0.24	0.30	0.50	0.50	0.29	0.28	0.58
1	14	0.54	0.80	0.81	0.60	0.57	0.55	0.60
1	21	0.71	0.90	0.85	0.60	0.61	0.58	0.65
2	2	0.16	0.25	0.26	0.10	0.35	0.63	0.15
2	10	0.40	0.60	0.60	0.50	0.39	0.53	0.20
2	12	0.16	0.05	0.15	0.00	0.35	0.43	0.16
2	18	0.30	0.40	0.40	0.40	0.33	0.48	0.10
3	1	0.07	0.50	0.55	0.20	0.35	0.18	0.29
3	3	0.24	0.30	0.45	0.40	0.34	0.13	0.23
3	10	0.07	0.10	0.25	0.30	0.17	0.10	0.03
3	16	0.19	0.50	0.61	0.55	0.32	0.15	0.37
4	1	0.57	0.65	0.55	0.40	0.49	0.53	0.60
4	4	0.43	0.40	0.55	0.60	0.52	0.23	0.28
4	11	0.39	0.55	0.15	0.30	0.48	0.00	0.20
4	12	0.21	0.40	0.25	0.30	0.37	0.30	0.08
4	14	0.67	0.80	0.71	0.50	0.60	0.68	0.63
5	1	0.74	0.85	0.90	0.40	0.35	0.75	0.51
5	4	0.56	0.45	0.80	0.40	0.37	0.43	0.39
5	13	0.13	0.50	0.45	0.70	0.51	0.78	0.50
5	14	0.30	0.60	0.40	0.40	0.74	0.50	0.34
5	17	0.20	0.40	0.39	0.40	0.37	0.53	0.50

Raw Scores for the 1999 SQA Private Service Sector

Ratee	Rater	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7
1	3	0.30	0.35	0.45	0.30	0.33	0.38	0.33
1	4	0.80	0.80	0.69	0.45	0.58	0.48	0.35
1	5	0.44	0.45	0.45	0.30	0.34	0.08	0.14
1	10	0.44	0.30	0.35	0.30	0.40	0.28	0.28
1	19	0.67	0.55	0.51	0.70	0.40	0.32	0.25
2	3	0.37	0.30	0.35	0.30	0.38	0.23	0.28
2	5	0.20	0.25	0.35	0.10	0.26	0.23	0.09
2	17	0.16	0.15	0.30	0.40	0.23	0.33	0.18
2	18	0.24	0.30	0.40	0.30	0.30	0.18	0.30
2	19	0.36	0.35	0.51	0.40	0.51	0.30	0.41
3	3	0.27	0.30	0.30	0.30	0.26	0.40	0.27
3	5	0.47	0.35	0.35	0.30	0.29	0.18	0.03
3	7	0.47	0.60	0.60	0.60	0.65	0.58	0.75
3	13	0.10	0.15	0.25	0.20	0.23	0.25	0.08

Raw Scores for the 1999 SQA Public Local Sector

Ratee	Rater	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7
1	2	0.50	0.25	0.20	0.00	0.17	0.23	0.06
1	6	0.34	0.40	0.45	0.40	0.39	0.33	0.48
1	7	0.57	0.65	0.70	0.60	0.68	0.78	0.66
1	8	0.33	0.30	0.35	0.30	0.34	0.13	0.31
1	9	0.31	0.30	0.59	0.10	0.24	0.25	0.51
1	19	0.54	0.45	0.65	0.30	0.57	0.60	0.43
2	2	0.20	0.15	0.25	0.10	0.17	0.18	0.00
2	3	0.47	0.60	0.50	0.50	0.55	0.43	0.21
2	13	0.43	0.55	0.35	0.30	0.55	0.43	0.00
2	16	0.46	0.53	0.63	0.75	0.50	0.38	0.38
2	20	0.24	0.30	0.15	0.10	0.26	0.20	0.08

Raw Scores for the 1999 SQA Public State and Federal Sector

Ratee	Rater	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7
1	6	0.31	0.10	0.25	0.10	0.23	0.30	0.14
1	8	0.43	0.50	0.45	0.50	0.46	0.38	0.25
1	9	0.47	0.60	0.27	0.40	0.33	0.35	0.36
1	15	0.40	0.45	0.45	0.20	0.39	0.38	0.08
1	16	0.71	0.55	0.59	0.50	0.63	0.38	0.52
2	6	0.40	0.25	0.51	0.40	0.37	0.33	0.34
2	8	0.50	0.45	0.55	0.50	0.53	0.48	0.55
2	11	0.41	0.35	0.70	0.50	0.31	0.30	0.11
2	15	0.49	0.60	0.55	0.40	0.36	0.38	0.19

APPENDIX B INTERRATER RELIABILITY CALCULATION

Private Manufacturing Sector, Dimension 1:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.3481	4	.0870		
Within People	.4028	15	.0269		
Between Measures	.0779	3	.0260	.9595	.4433
Residual	.3249	12	.0271		
Total	.7510	19	.0395		
Grand Mean	.3310				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .3590

95.00% C.I.: Lower = -.0385 Upper = .8712

F = 3.2406 DF = (4, 15.0) Sig. = .0419 (Test Value = .0000)

Average Measure Intraclass Correlation = .6914

95.00% C.I.: Lower = -.1739 Upper = .9644

F = 3.2406 DF = (4, 15.0) Sig. = .0419 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .6889

Private Manufacturing Sector, Dimension 2:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.3380	4	.0845		
Within People	.6225	15	.0415		
Between Measures	.1285	3	.0428	1.0405	.4099
Residual	.4940	12	.0412		
Total	.9605	19	.0506		
Grand Mean	.4850				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .2057

95.00% C.I.: Lower = -.1315 Upper = .8061

F = 2.0361 DF = (4, 15.0) Sig. = .1407 (Test Value = .0000)

Average Measure Intraclass Correlation = .5089

95.00% C.I.: Lower = -.8684 Upper = .9433

F = 2.0361 DF = (4, 15.0) Sig. = .1407 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .5128

Private Manufacturing Sector, Dimension 3:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.3181	4	.0795		
Within People	.6772	15	.0451		
Between Measures	.1295	3	.0432	.9460	.4491
Residual	.5477	12	.0456		
Total	.9954	19	.0524		
Grand Mean	.5090				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .1599

95.00% C.I.: Lower = -.1551 Upper = .7808

F = 1.7615 DF = (4, 15.0) Sig. = .1891 (Test Value = .0000)

Average Measure Intraclass Correlation = .4323

95.00% C.I.: Lower = -1.1597 Upper = .9344

F = 1.7615 DF = (4, 15.0) Sig. = .1891 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .4261

Private Manufacturing Sector, Dimension 4:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.1875	4	.0469		
Within People	.3819	15	.0255		
Between Measures	.2174	3	.0725	5.2857	.0149
Residual	.1645	12	.0137		
Total	.5694	19	.0300		
Grand Mean	.4125				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .1738

95.00% C.I.: Lower = -.1481 Upper = .7888

F = 1.8412 DF = (4, 15.0) Sig. = .1735 (Test Value = .0000)

Average Measure Intraclass Correlation = .4569

95.00% C.I.: Lower = -1.0661 Upper = .9373

F = 1.8412 DF = (4, 15.0) Sig. = .1735 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .7076

Private Manufacturing Sector, Dimension 5:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.1156	4	.0289		
Within People	.2581	15	.0172		
Between Measures	.0839	3	.0280	1.9280	.1789
Residual	.1741	12	.0145		
Total	.3737	19	.0197		
Grand Mean	.4105				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .1453

95.00% C.I.: Lower = -.1622 Upper = .7720

F = 1.6800 DF = (4, 15.0) Sig. = .2067 (Test Value = .0000)

Average Measure Intraclass Correlation = .4048

95.00% C.I.: Lower = -1.2644 Upper = .9312

F = 1.6800 DF = (4, 15.0) Sig. = .2067 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0 N of Items = 4

Alpha = .4980

Private Manufacturing Sector, Dimension 6:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.5446	4	.1361		
Within People	.2762	15	.0184		
Between Measures	.0593	3	.0198	1.0933	.3895
Residual	.2169	12	.0181		
Total	.8208	19	.0432		
Grand Mean	.4370				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .6151

95.00% C.I.: Lower = .1908 Upper = .9403

F = 7.3924 DF = (4, 15.0) Sig. = .0017 (Test Value = .0000)

Average Measure Intraclass Correlation = .8647

95.00% C.I.: Lower = .4854 Upper = .9844

F = 7.3924 DF = (4, 15.0) Sig. = .0017 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .8672

Private Manufacturing Sector, Dimension 7:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.4475	4	.1119		
Within People	.3306	15	.0220		
Between Measures	.1153	3	.0384	2.1406	.1483
Residual	.2154	12	.0179		
Total	.7781	19	.0410		
Grand Mean	.3605				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .5047

95.00% C.I.: Lower = .0771 Upper = .9148

F = 5.0753 DF = (4, 15.0) Sig. = .0087 (Test Value = .0000)

Average Measure Intraclass Correlation = .8030

95.00% C.I.: Lower = .2504 Upper = .9772

F = 5.0753 DF = (4, 15.0) Sig. = .0087 (Test Value = .0000)

Reliability Coefficients

N of Cases = 5.0

N of Items = 4

Alpha = .8396

Private Service Sector, Dimension 1:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0010	1	.0010		
Within People	.0963	6	.0160		
Between Measures	.0685	3	.0228	2.4709	.2385
Residual	.0277	3	.0092		
Total	.0973	7	.0139		
Grand Mean	.4613				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.3059

95.00% C.I.: Lower = -.3302 Upper = .9356

F = .0631 DF = (1, 6.0) Sig. = .8100 (Test Value = .0000)

Average Measure Intraclass Correlation = -14.8477

95.00% C.I.: Lower = -138.668 Upper = .9831

F = .0631 DF = (1, 6.0) Sig. = .8100 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = -.81317

Private Service Sector, Dimension 2:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0003	1	.0003		
Within People	.2194	6	.0366		
Between Measures	.1909	3	.0636	6.7143	.0761
Residual	.0284	3	.0095		
Total	.2197	7	.0314		
Grand Mean	.4188				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.3295

95.00% C.I.: Lower = -.3329 Upper = .6367

F = .0085 DF = (1, 6.0) Sig. = .9294 (Test Value = .0000)

Average Measure Intraclass Correlation = -116.000

95.00% C.I.: Lower = -1030.13 Upper = .8751

F = .0085 DF = (1, 6.0) Sig. = .9294 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = -.293333

Private Service Sector, Dimension 3:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0703	1	.0703		
Within People	.0987	6	.0164		
Between Measures	.0369	3	.0123	.5983	.6583
Residual	.0617	3	.0206		
Total	.1690	7	.0241		
Grand Mean	.4838				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .4502

95.00% C.I.: Lower = -.1477 Upper = .9990

F = 4.2754 DF = (1, 6.0) Sig. = .0842 (Test Value = .0000)

Average Measure Intraclass Correlation = .7661

95.00% C.I.: Lower = -1.0614 Upper = .9998

F = 4.2754 DF = (1, 6.0) Sig. = .0842 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = .7073

Private Service Sector, Dimension 4:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0050	1	.0050		
Within People	.0700	6	.0117		
Between Measures	.0150	3	.0050	.2727	.8429
Residual	.0550	3	.0183		
Total	.0750	7	.0107		
Grand Mean	.4250				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.1667

95.00% C.I.: Lower = -.3121 Upper = .9901

F = .4286 DF = (1, 6.0) Sig. = .5370 (Test Value = .0000)

Average Measure Intraclass Correlation = -1.3333

95.00% C.I.: Lower = -19.5639 Upper = .9975

F = .4286 DF = (1, 6.0) Sig. = .5370 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = -2.6667

Private Service Sector, Dimension 5:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0024	1	.0024		
Within People	.1098	6	.0183		
Between Measures	.0603	3	.0201	1.2194	.4372
Residual	.0495	3	.0165		
Total	.1122	7	.0160		
Grand Mean	.4100				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.2763

95.00% C.I.: Lower = -.3266 Upper = .9689

F = .1339 DF = (1, 6.0) Sig. = .7269 (Test Value = .0000)

Average Measure Intraclass Correlation = -6.4660

95.00% C.I.: Lower = -64.7985 Upper = .9920

F = .1339 DF = (1, 6.0) Sig. = .7269 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = -.57279

Private Service Sector, Dimension 6:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0008	1	.0008		
Within People	.0230	6	.0038		
Between Measures	.0171	3	.0057	2.8898	.2033
Residual	.0059	3	.0020		
Total	.0238	7	.0034		
Grand Mean	.3625				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.2464

95.00% C.I.: Lower = -.3229 Upper = .9799

F = .2092 DF = (1, 6.0) Sig. = .6635 (Test Value = .0000)

Average Measure Intraclass Correlation = -3.7812

95.00% C.I.: Lower = -41.1376 Upper = .9949

F = .2092 DF = (1, 6.0) Sig. = .6635 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 4

Alpha = -1.4583

Private Service Sector, Dimension 7:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0008	1	.0008		
Within People	.1912	6	.0319		
Between Measures	.0413	3	.0138	.2752	.8413
Residual	.1499	3	.0500		
Total	.1920	7	.0274		
Grand Mean	.3075				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.3223

95.00% C.I.: Lower = -.3321 Upper = .8492

F = .0251 DF = (1, 6.0) Sig. = .8793 (Test Value = .0000)

Average Measure Intraclass Correlation = -.388229

95.00% C.I.: Lower = -.349.963 Upper = .9575

F = .0251 DF = (1, 6.0) Sig. = .8793 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0 N of Items = 4

Alpha = -.614583

Public Local Agencies Sector, Dimension 1:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.2053	2	.1026		
Within People	.2720	9	.0302		
Between Measures	.0565	3	.0188	.5242	.6815
Residual	.2155	6	.0359		
Total	.4773	11	.0434		
Grand Mean	.3742				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .3746

95.00% C.I.: Lower = -.1129 Upper = .9707

F = 3.3956 DF = (2, 9.0) Sig. = .0797 (Test Value = .0000)

Average Measure Intraclass Correlation = .7055

95.00% C.I.: Lower = -.6830 Upper = .9925

F = 3.3956 DF = (2, 9.0) Sig. = .0797 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = .6500

Public Local Agencies Sector, Dimension 2:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.1050	2	.0525		
Within People	.2825	9	.0314		
Between Measures	.0242	3	.0081	.1871	.9014
Residual	.2583	6	.0431		
Total	.3875	11	.0352		
Grand Mean	.3750				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .1439

95.00% C.I.: Lower = -.2148 Upper = .9419

F = 1.6726 DF = (2, 9.0) Sig. = .2412 (Test Value = .0000)

Average Measure Intraclass Correlation = .4021

95.00% C.I.: Lower = -2.4167 Upper = .9848

F = 1.6726 DF = (2, 9.0) Sig. = .2412 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = .1799

Public Local Agencies Sector, Dimension 3:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0546	2	.0273		
Within People	.1354	9	.0150		
Between Measures	.0215	3	.0072	.3782	.7725
Residual	.1139	6	.0190		
Total	.1900	11	.0173		
Grand Mean	.4300				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .1692

95.00% C.I.: Lower = -.2057 Upper = .9463

F = 1.8146 DF = (2, 9.0) Sig. = .2177 (Test Value = .0000)

Average Measure Intraclass Correlation = .4489

95.00% C.I.: Lower = -.21493 Upper = .9860

F = 1.8146 DF = (2, 9.0) Sig. = .2177 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = .3048

Public Local Agencies Sector, Dimension 4:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0388	2	.0194		
Within People	.2569	9	.0285		
Between Measures	.0706	3	.0235	.7584	.5569
Residual	.1862	6	.0310		
Total	.2956	11	.0269		
Grand Mean	.3625				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.0873

95.00% C.I.: Lower = -.2825 Upper = .8655

F = .6788 DF = (2, 9.0) Sig. = .5314 (Test Value = .0000)

Average Measure Intraclass Correlation = -.4731

95.00% C.I.: Lower = -.74184 Upper = .9626

F = .6788 DF = (2, 9.0) Sig. = .5314 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = -.6022

Public Local Agencies Sector, Dimension 5:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0168	2	.0084		
Within People	.1972	9	.0219		
Between Measures	.0294	3	.0098	.3508	.7905
Residual	.1678	6	.0280		
Total	.2140	11	.0195		
Grand Mean	.3775				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.1818

95.00% C.I.: Lower = -.3041 Upper = .7796

F = .3846 DF = (2, 9.0) Sig. = .6914 (Test Value = .0000)

Average Measure Intraclass Correlation = -1.6004

95.00% C.I.: Lower = -13.8605 Upper = .9340

F = .3846 DF = (2, 9.0) Sig. = .6914 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = -.23185

Public Local Agencies Sector, Dimension 6:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0277	2	.0138		
Within People	.1828	9	.0203		
Between Measures	.0769	3	.0256	1.4524	.3183
Residual	.1059	6	.0176		
Total	.2104	11	.0191		
Grand Mean	.2925				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.0867

95.00% C.I.: Lower = -.2824 Upper = .8658

F = .6808 DF = (2, 9.0) Sig. = .5305 (Test Value = .0000)

Average Measure Intraclass Correlation = -.4690

95.00% C.I.: Lower = -.73947 Upper = .9627

F = .6808 DF = (2, 9.0) Sig. = .5305 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = -.2765

Public Local Agencies Sector, Dimension 7:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0039	2	.0020		
Within People	.3695	9	.0411		
Between Measures	.1636	3	.0545	1.5893	.2876
Residual	.2059	6	.0343		
Total	.3735	11	.0340		
Grand Mean	.2750				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.3123

95.00% C.I.: Lower = -.3296 Upper = .1828

F = .0481 DF = (2, 9.0) Sig. = .9533 (Test Value = .0000)

Average Measure Intraclass Correlation = -.19.7904

95.00% C.I.: Lower = -.117.811 Upper = .4721

F = .0481 DF = (2, 9.0) Sig. = .9533 (Test Value = .0000)

Reliability Coefficients

N of Cases = 3.0

N of Items = 4

Alpha = -.16.3769

Public State and Federal Agencies Sector, Dimension 1:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0212	1	.0212		
Within People	.1237	8	.0155		
Between Measures	.0253	4	.0063	.2577	.8913
Residual	.0983	4	.0246		
Total	.1448	9	.0161		
Grand Mean	.4060				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .0687

95.00% C.I.: Lower = -.1959 Upper = .9962

F = 1.3687 DF = (1, 8.0) Sig. = .2757 (Test Value = .0000)

Average Measure Intraclass Correlation = .2694

95.00% C.I.: Lower = -4.5315 Upper = .9992

F = 1.3687 DF = (1, 8.0) Sig. = .2757 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0 N of Items = 5

Alpha = -.1619

Public State and Federal Agencies Sector, Dimension 2:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0032	1	.0032		
Within People	.2555	8	.0319		
Between Measures	.1836	4	.0459	2.5509	.1933
Residual	.0720	4	.0180		
Total	.2588	9	.0288		
Grand Mean	.4080				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.2191

95.00% C.I.: Lower = -.2458 Upper = .9505

F = .1014 DF = (1, 8.0) Sig. = .7583 (Test Value = .0000)

Average Measure Intraclass Correlation = -.8580

95.00% C.I.: Lower = -.73.6339 Upper = .9897

F = .1014 DF = (1, 8.0) Sig. = .7583 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0 N of Items = 5

Alpha = -.45525

Public State and Federal Agencies Sector, Dimension 3:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0504	1	.0504		
Within People	.3090	8	.0386		
Between Measures	.1699	4	.0425	1.2208	.4257
Residual	.1391	4	.0348		
Total	.3594	9	.0399		
Grand Mean	.4470				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .0575

95.00% C.I.: Lower = -.1984 Upper = .9960

F = 1.3051 DF = (1, 8.0) Sig. = .2863 (Test Value = .0000)

Average Measure Intraclass Correlation = .2338

95.00% C.I.: Lower = -4.8009 Upper = .9992

F = 1.3051 DF = (1, 8.0) Sig. = .2863 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 5

Alpha = .3100

Public State and Federal Agencies Sector, Dimension 4:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0122	1	.0122		
Within People	.5380	8	.0672		
Between Measures	.3940	4	.0985	2.7361	.1766
Residual	.1440	4	.0360		
Total	.5502	9	.0611		
Grand Mean	.3150				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.1956

95.00% C.I.: Lower = -.2425 Upper = .9720

F = .1822 DF = (1, 8.0) Sig. = .6808 (Test Value = .0000)

Average Measure Intraclass Correlation = -4.4898

95.00% C.I.: Lower = -40.5626 Upper = .9943

F = .1822 DF = (1, 8.0) Sig. = .6808 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 5

Alpha = -1.9388

Public State and Federal Agencies Sector, Dimension 5:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	9.0E-05	1	.0001		
Within People	.3167	8	.0396		
Between Measures	.2045	4	.0511	1.8213	.2878
Residual	.1123	4	.0281		
Total	.3168	9	.0352		
Grand Mean	.4030				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.2493

95.00% C.I.: Lower = -.2499 Upper = .1903

F = .0023 DF = (1, 8.0) Sig. = .9631 (Test Value = .0000)

Average Measure Intraclass Correlation = -.438889

95.00% C.I.: Lower = -.3329.35 Upper = .5402

F = .0023 DF = (1, 8.0) Sig. = .9631 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 5

Alpha = -.310.833

Public State and Federal Agencies Sector, Dimension 6:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.0203	1	.0203		
Within People	.3518	8	.0440		
Between Measures	.1933	4	.0483	1.2198	.4260
Residual	.1585	4	.0396		
Total	.3721	9	.0413		
Grand Mean	.3690				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = -.1210

95.00% C.I.: Lower = -.2313 Upper = .9888

F = .4604 DF = (1, 8.0) Sig. = .5166 (Test Value = .0000)

Average Measure Intraclass Correlation = -1.1719

95.00% C.I.: Lower = -15.4428 Upper = .9977

F = .4604 DF = (1, 8.0) Sig. = .5166 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 5

Alpha = -.9568

Public State and Federal Agencies Sector, Dimension 7:

RELIABILITY ANALYSIS - SCALE (ALPHA)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	.1690	1	.1690		
Within People	.3092	8	.0387		
Between Measures	.2574	4	.0644	4.9699	.0748
Residual	.0518	4	.0129		
Total	.4782	9	.0531		
Grand Mean	.2640				

Intraclass Correlation Coefficient

One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = .4028

95.00% C.I.: Lower = -.0923 Upper = .9988

F = 4.3720 DF = (1, 8.0) Sig. = .0699 (Test Value = .0000)

Average Measure Intraclass Correlation = .7713

95.00% C.I.: Lower = -.7317 Upper = .9998

F = 4.3720 DF = (1, 8.0) Sig. = .0699 (Test Value = .0000)

Reliability Coefficients

N of Cases = 2.0

N of Items = 5

Alpha = .9234

APPENDIX C HALO ERROR CALCULATION

Private Manufacturing Sector: Examiner 1

Correlations

	DIM_1	DIM_2	DIM_3	DIM_4	DIM_5	DIM_6	DIM_7
DIM_1 Pearson Correlation	1.000	.908	.650	.920	.156	.970*	.838
Sig. (2-tailed)	.	.092	.350	.080	.844	.030	.162
N	4	4	4	4	4	4	4
DIM_2 Pearson Correlation	.908	1.000	.906	.679	.128	.977*	.646
Sig. (2-tailed)	.092	.	.094	.321	.872	.023	.354
N	4	4	4	4	4	4	4
DIM_3 Pearson Correlation	.650	.906	1.000	.333	-.031	.796	.288
Sig. (2-tailed)	.350	.094	.	.667	.969	.204	.712
N	4	4	4	4	4	4	4
DIM_4 Pearson Correlation	.920	.679	.333	1.000	.031	.796	.813
Sig. (2-tailed)	.080	.321	.667	.	.969	.204	.187
N	4	4	4	4	4	4	4
DIM_5 Pearson Correlation	.156	.128	-.031	.031	1.000	.222	.606
Sig. (2-tailed)	.844	.872	.969	.969	.	.778	.394
N	4	4	4	4	4	4	4
DIM_6 Pearson Correlation	.970*	.977*	.796	.796	.222	1.000	.787
Sig. (2-tailed)	.030	.023	.204	.204	.778	.	.213
N	4	4	4	4	4	4	4
DIM_7 Pearson Correlation	.838	.646	.288	.813	.606	.787	1.000
Sig. (2-tailed)	.162	.354	.712	.187	.394	.213	.
N	4	4	4	4	4	4	4
N	4	4	4	4	4	4	4

* Correlation is significant at the 0.05 level (2-tailed).

The determinant of the dimension intercorrelation matrix is 7.758E-14.

Private Manufacturing Sector: Examiner 2

Correlations

	DIM_1	DIM_2	DIM_3	DIM_4	DIM_5	DIM_6	DIM_7
DIM_1 Pearson Correlation	1.000	.938	.831	.639	-.958	.973	.917
Sig. (2-tailed)		.225	.375	.558	.184	.149	.262
N	3	3	3	3	3	3	3
DIM_2 Pearson Correlation	.938	1.000	.972	.866	-.998*	.832	.998*
Sig. (2-tailed)	.225		.150	.333	.041	.374	.037
N	3	3	3	3	3	3	3
DIM_3 Pearson Correlation	.831	.972	1.000	.959	-.955	.679	.984
Sig. (2-tailed)	.375	.150		.183	.191	.524	.114
N	3	3	3	3	3	3	3
DIM_4 Pearson Correlation	.639	.866	.959	1.000	-.832	.444	.893
Sig. (2-tailed)	.558	.333	.183		.374	.707	.297
N	3	3	3	3	3	3	3
DIM_5 Pearson Correlation	-.958	-.998*	-.955	-.832	1.000	-.866	-.993
Sig. (2-tailed)	.184	.041	.191	.374		.333	.077
N	3	3	3	3	3	3	3
DIM_6 Pearson Correlation	.973	.832	.679	.444	-.866	1.000	.799
Sig. (2-tailed)	.149	.374	.524	.707	.333		.411
N	3	3	3	3	3	3	3
DIM_7 Pearson Correlation	.917	.998*	.984	.893	-.993	.799	1.000
Sig. (2-tailed)	.262	.037	.114	.297	.077	.411	
N	3	3	3	3	3	3	3

* Correlation is significant at the 0.05 level (2-tailed).

The determinant of the dimension intercorrelation matrix is 1.185E-16.

Public Local Sector: Examiner 1

Correlations

		DIM_1	DIM_2	DIM_3	DIM_4	DIM_5	DIM_6	DIM_7
DIM_1	Pearson Correlation	1.000	-.225	.106	.	.948	-.982	-.071
	Sig. (2-tailed)	.	.856	.932	.	.206	.120	.955
	N	3	3	3	3	3	3	3
DIM_2	Pearson Correlation	-.225	1.000	.945	.	.096	.404	.988
	Sig. (2-tailed)	.856	.	.212	.	.939	.735	.099
	N	3	3	3	3	3	3	3
DIM_3	Pearson Correlation	.106	.945	1.000	.	.416	.082	.984
	Sig. (2-tailed)	.932	.212	.	.	.727	.948	.113
	N	3	3	3	3	3	3	3
DIM_4	Pearson Correlation	.	.	.	.	.	.	.
	Sig. (2-tailed)	.	.	.	.	.	.	.
	N	3	3	3	3	3	3	3
DIM_5	Pearson Correlation	.948	.096	.416	.	1.000	-.872	.249
	Sig. (2-tailed)	.206	.939	.727	.	.	.326	.840
	N	3	3	3	3	3	3	3
DIM_6	Pearson Correlation	-.982	.404	.082	.	-.872	1.000	.257
	Sig. (2-tailed)	.120	.735	.948	.	.326	.	.835
	N	3	3	3	3	3	3	3
DIM_7	Pearson Correlation	-.071	.988	.984	.	.249	.257	1.000
	Sig. (2-tailed)	.955	.099	.113	.	.840	.835	.
	N	3	3	3	3	3	3	3

a. Cannot be computed because at least one of the variables is constant.

The determinant of the dimension intercorrelation matrix can not be calculated.

Public Local Sector: Examiner 2

Correlations

		DIM_1	DIM_2	DIM_3	DIM_4	DIM_5	DIM_6	DIM_7
DIM_1	Pearson Correlation	1.000	.811	.410	.995	.719	-.686	-.153
	Sig. (2-tailed)	.	.398	.731	.065	.489	.519	.902
	N	3	3	3	3	3	3	3
DIM_2	Pearson Correlation	.811	1.000	.866	.866	.990	-.982	.454
	Sig. (2-tailed)	.398	.	.333	.333	.091	.121	.700
	N	3	3	3	3	3	3	3
DIM_3	Pearson Correlation	.410	.866	1.000	.500	.929	-.945	.839
	Sig. (2-tailed)	.731	.333	.	.667	.242	.212	.367
	N	3	3	3	3	3	3	3
DIM_4	Pearson Correlation	.995	.866	.500	1.000	.786	-.756	-.052
	Sig. (2-tailed)	.065	.333	.667	.	.425	.454	.967
	N	3	3	3	3	3	3	3
DIM_5	Pearson Correlation	.719	.990	.929	.786	1.000	-.999*	.577
	Sig. (2-tailed)	.489	.091	.242	.425	.	.030	.609
	N	3	3	3	3	3	3	3
DIM_6	Pearson Correlation	-.686	-.982	-.945	-.756	-.999*	1.000	-.614
	Sig. (2-tailed)	.519	.121	.212	.454	.030	.	.579
	N	3	3	3	3	3	3	3
DIM_7	Pearson Correlation	-.153	.454	.839	-.052	.577	-.614	1.000
	Sig. (2-tailed)	.902	.700	.367	.967	.609	.579	.
	N	3	3	3	3	3	3	3

* Correlation is significant at the 0.05 level (2-tailed).

The determinant of the dimension intercorrelation matrix is -6.856E-18.

VITA

Jianming Shen was born in Jiangsu province of China on November 30, 1967. He graduated from Shanghai Jiaotong University and earned a Bachelors of Science Degree in Mechanical Engineering in 1990. He then attended Shanghai Academy of Spaceflight Technology where he earned a Master of Science Degree in Mechanical Engineering in 1993. In May of 1993 he became an employee of Shanghai Spaceflight Power Machinery Institute. In August of 2001 he completed his Master of Science Degree in Industrial Engineering at The University of Tennessee.