1998

Multi-rater feedback assessment : an analysis of a feedback survey

Dana C. Larsen
San Jose State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Multi-Rater Feedback Assessment: An Analysis of a Feedback Survey

A Thesis

Presented to

the Faculty of the Department of Psychology

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

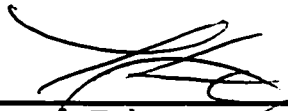
Master of Science

by

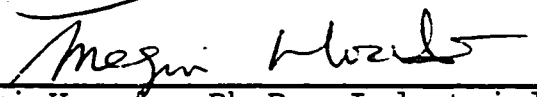
Dana C. Larsen

December, 1998

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY



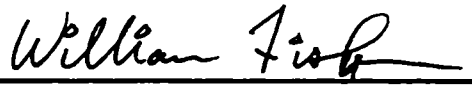
Howard Tokunaga, Ph.D., Industrial/Organizational Psychology



Megumi Hosoda, Ph.D., Industrial/Organizational Psychology

Yoav Harlev, M.A., Organizational Effectiveness Manager

APPROVED FOR THE UNIVERSITY



ABSTRACT

MULTI-RATER FEEDBACK ASSESSMENT: AN ANALYSIS OF A FEEDBACK SURVEY

by Dana C. Larsen

The purpose of this study was to analyze the factor structure and reliability of a 34-item multi-rater assessment instrument used by a high technology firm for leadership development. Using data from 750 survey participants collected over 3 years, a principal components analysis was conducted to define the underlying constructs being measured by the instrument. Cronbach's alpha was generated and analyzed to determine the relative contribution of individual items to scale reliability for the instrument as a whole. In addition, reliability coefficients were computed to assess item contribution to scale reliability for each rater category (self-raters, colleague-raters, direct report raters, and supervisor-raters). Recommendations for instrument revision are discussed.

ACKNOWLEDGMENTS

Many individuals contributed in one way or another to the completion of this thesis study. I would like to, first and foremost, thank my wife Lisa for her long-suffering support. Also, special thanks are due to the members of my thesis committee: Dr. Howard Tokunaga for his encouragement and psychometric expertise and guidance, to Dr. Megumi Hosoda for her critical eye and drive for improvement, and to Yoav Harlev, without whose instrument and desire for its improvement this research would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

Multi-Rater Feedback Assessment: An Analysis of a Feedback Survey

Running Head: MULTI-RATER ASSESSMENT

Abstract

The purpose of this study was to analyze the factor structure and reliability of a 34-item multi-rater assessment instrument used by a high technology firm for leadership development. Using data from 750 survey participants collected over 3 years, a principal components analysis was conducted to define the underlying constructs being measured by the instrument. Cronbach's alpha was generated and analyzed to determine the relative contribution of individual items to scale reliability for the instrument as a whole. In addition, reliability coefficients were computed to assess item contribution to scale reliability for each rater category (self-raters, colleague-raters, direct report raters, and supervisor-raters). Recommendations for instrument revision are discussed.

Multi-Rater Feedback Assessment: An Analysis of a Feedback Survey

During the 1980's companies began experimenting with performance assessment systems that did not conform to traditional single-rater systems in which managers were the sole judge of performance. Multi-rater assessment systems, in which an employee is evaluated from many different perspectives (e.g. self, supervisor, peers, and subordinates), promised to provide individuals with feedback that was more reliable and less subject to biases inherent in single-rater assessment methodologies. Also, this new form of assessment would provide detailed performance feedback from a variety of perspectives, allowing the rated individual to identify patterns of behavior perceived as "good" and those needing improvement. The purpose of this study is to analyze the underlying constructs and psychometric properties of a multi-rater feedback instrument used within a division of a high technology firm.

Multi-rater Assessment Defined

Multi-rater assessment was defined by London and Smither (1995) as the process by which performance evaluations of an employee are collected from more than one source. Within the realm of multi-rater feedback, however, varying configurations are often given different names to better describe the nature of the rating relationship. 360-degree feedback, for example, may include supervisors,

subordinates, peers, self-ratings, customers, and suppliers from both within and outside of the immediate organization (Yammarino & Atwater, 1993). Upward feedback, on the other hand, involves the collection of ratings from multiple subordinates and the comparison of those ratings to self-ratings (Reilly, Smither, & Vasilopoulos, 1996; Atwater, Roush, & Fischthal, 1995; Smither, London, Reilly, & Millsap, 1995). Other possible feedback sources include customers both internal and external to the organization. Because data are not always collected from the entire "360 degrees", the term multi-rater feedback will be used in this paper to represent any assessment system in which performance feedback information is collected from more than one source.

Current Use of Multi-rater Assessment

Edwards and Ewen (1996) report that 90% of Fortune 1000 companies use some form of multi-rater assessment. Bohl (1996) found that 13% of companies surveyed used a full "360 degree" multi-rater system either administered internal to the company or purchased from an external vendor, while another 9% used a "partial" multi-rater system (e.g. upward feedback or customer feedback). In addition, 90% of the companies who used multi-rater assessment, used it to make personnel decisions (e.g., merit increases, promotions, terminations) as a part of their standard performance appraisal process, while the remainder utilized it primarily

as a developmental tool (e.g., improvement of interpersonal, managerial and technical skills). Overall, multi-rater feedback systems were perceived by individuals surveyed to provide more valid, consistent, and fair results.

Benefits of Multi-rater Assessment

A review of the literature suggests that there are some benefits to using a multi-rater feedback assessment. These benefits include improvements to the assessment process, performance improvements, and improvements to measurement accuracy.

Process improvements. Wanguri (1995), in a review of performance appraisal literature lists some of the benefits inherent in multi-rater feedback systems. First, multi-rater assessments generally increase participation in the performance management process. Also, multi-rater assessments are perceived as being more accurate and fair than traditional performance reviews. This is due to the fact that individuals being rated often select those who will rate them and also complete a self-assessment. When used for making selection decisions for promotional opportunities, multi-rater assessments provide more quality information about potential candidates due to the increased number of perspectives being measured, limiting single rater biases. Also, when efficiently administered, multi-rater feedback assessments decrease the time taken to complete the

performance measurement process, and are more cost effective.

Bohl (1996) found in a survey of 756 companies that those using multi-rater feedback systems perceived benefits differently from those using traditional assessment methodologies. Eighty-two percent of individuals surveyed from companies using multi-rater feedback systems felt that their systems improved employee understanding and self-awareness, compared to only 70% of those from companies using a traditional performance appraisal system. It may be that multi-rater assessments provide better understanding by allowing employees to compare their self-perceptions with the perceptions of others rather than only receiving input from their immediate supervisor. Also, 83% of those using multi-rater assessment felt that their assessment methodologies had improved communication between supervisors and their staffs, while only 66% of respondents using traditional assessment systems felt like communication had improved. This result may be due to the fact that the anonymous nature of multi-rater assessments provides raters the opportunity to give more open and honest assessments of a ratee's performance without fear of reprisal.

Performance improvements. In the survey reported by Bohl (1996), 55% of those using traditional performance assessment systems felt that their systems promoted better performance. An additional 45% of these respondents

disagreed that their traditional assessment system made any contribution to an improvement in on-the-job performance. In contrast, 68% of those using multi-rater feedback perceived that performance generally improved following feedback. Others found that second round subordinate and colleague ratings of managers increased following multi-rater assessment, but supervisor and customer ratings did not (Bernardin, Hagan, & Kane's study as cited in Reilly, Smither, & Vasilopoulos, 1996).

Bernardin et al. (1995) suggested that because colleagues and subordinates generally provided the most negative feedback, their input inspired greater levels of motivation for goal setting and behavior change. This contention is supported by a study conducted by Atwater, Roush, and Fischthal (1995), in which each of 978 student leaders from the U.S. Naval Academy were rated by the subordinates in their squad (N=4). A total of 1232 raters participated. Self and subordinate ratings of leadership performance were collected three times during the academic year. The researchers found that ratings of student leaders improved after leaders received feedback from subordinates, and those leaders receiving particularly negative feedback improved the most. In addition, leaders whose self-ratings were initially high and who received negative feedback were more likely to rate themselves lower in subsequent ratings while those leaders rating themselves low initially

increased their self-ratings following positive feedback from subordinates. These results suggest that multi-rater feedback assessments may motivate changes in leadership behavior especially when the feedback is negative. Additionally, feedback may bring a leader's self perceptions in line with those with whom he or she works.

Smither, London, Reilly, and Millsap (1995) performed a longitudinal study of managerial performance when upward feedback (e.g. employees rating their immediate supervisors) was the primary performance assessment tool. Managers within a large corporation (N=238) were rated on 33 behavioral statements grouped into eight categories (i.e., coaching/support, commitment to quality/customer satisfaction, communication, creating a team environment, fairness, integrity and respect, participation/empowerment, and providing feedback). They found that managers whose initial level of performance was low in any one category improved their performance within that category substantially during the 6 month interval between administrations of the feedback instrument while those whose performance was rated higher showed little or no improvement.

In a longitudinal study of multi-rater assessment impact on performance improvement, 92 managers were rated by one or more subordinates four times over a period of two and a half years. Reilly, Smither, and Vasilopoulos (1996) found

that managers whose initial level of performance was rated low or moderate improved modestly, while those whose initial level of performance was rated high did not change significantly following the receipt of feedback. In addition, observed performance improvements were sustained and slightly enhanced over time.

Measurement accuracy. Thomas and Bretz (1994) suggest that getting input from multiple sources "reduces the sampling error by increasing the number of observations and reduces the effect of possible idiosyncratic biases" (p. 34). In addition, measuring a number of perspectives instead of one or two as in traditional performance evaluations increases the potential for variability of response and therefore the likelihood that responses will be reliable over time. Ideally this characteristic of multi-rater assessment would allow for reliable measurement using a relatively short instrument (Murphy & Davidshofer, 1997).

Drawbacks to Using Multi-Rater Feedback Systems

While some research has demonstrated advantages to collecting performance feedback from multiple raters, other researchers and practitioners have demonstrated the drawbacks associated with the use of multi-rater feedback assessment. These include logistical complexity, and some issues related to rating reliability and accuracy of measurement.

Logistical issues. Multi-rater systems, while tending to yield more meaningful results for the rated individual, are logistically intensive to operate efficiently. With traditional performance measurement systems, one individual was responsible for providing input into another's performance. With multi-rater assessment, however, multiple inputs must be collected, combined, and reported back in order to make the system effective. The result is that resources must be specifically dedicated to the maintenance of the program.

Lepsinger and Lucia (1997) provide a good review of the "extra-steps" involved in establishing and maintaining an effective multi-rater system. These steps include the selection of raters by category (e.g. colleagues, subordinates, customers, etc.), ensuring the anonymity of ratings, program orientation and feedback meetings, the administration of survey instruments (often in multiple forms), and report preparation. They also note that where external vendors are concerned, multi-rater feedback systems are costly compared to programs administered internal to an organization.

Rating reliability and accuracy. Feedback instruments themselves are subject to the same problems inherent in any measurement tool. The development and use of feedback instruments for multi-rater assessments must conform to standards of measurement consistency and accuracy associated

with the constructs being measured (DeVellis, 1993). An instrument's consistency of measurement may be established statistically through the estimation of reliability. Also, its tendency to measure the desired construct(s) may be accomplished by the estimation of validity. Because an instrument's scales must be reliable in order for measurement to be valid, a scale's reliability should be of primary concern.

In addition to measurement reliability, research has demonstrated that self-ratings have a tendency to be inflated relative to the ratings of others. Yammarino and Atwater (1993) report that this tendency for self-rater leniency makes self-ratings unreliable and suspect when compared to the ratings of others. This leniency bias has been demonstrated in a number of studies. Researchers studying self-perception of classroom performance found that students gave more accurate self-ratings of their classroom performance when they thought that the assessments were being used for research purposes than when they thought the same assessments were being used to determine grades (Farh & Werbel's study as cited in Nilsen & Campbell, 1993). Aamodt (1991) reports that self-ratings of teacher performance only moderately correlate (r=.29) with actual performance and that self-raters tended to inflate their ratings, especially when the ratings were being used for promotions or raises.

These studies demonstrate that self-raters are likely to inflate ratings of self-performance, particularly when there is a reward to be achieved. This has important implications for multi-rater assessment, especially when it is used to make performance management decisions (e.g. promotions, raises, etc.) and not solely for professional development. Nilsen and Campbell (1993) suggest that individual differences in self-perception accuracy do exist and that these inaccuracies demonstrate a problem inherent to the collection of self-ratings. However, they contend that although there are discrepancies between self-ratings and other ratings, those discrepancies remain stable over-time and may therefore be considered reliable self-perceptions of performance. Results reported by Wohlers, Hall, and London (1993) support this contention. They studied the question of self-subordinate agreement with employees representing different demographic groups (e.g. age, race and gender) and working for organizations of different types (i.e. public or private). The study found that individuals tend to maintain a favorable view of themselves which leads to inflated self-ratings. However, the inflation was consistent and deemed reliable.

Again, these results have implications for multi-rater assessment. If self-ratings, though inflated relative to other-ratings, represent accurate self-perceptions of performance, self-ratings remain a valuable facet of multi-

rater assessment, providing participants with the opportunity to compare and evaluate self-perceptions and the perceptions of others. The "self-other" distinction has been the subject of considerable debate among researchers, yet few studies have been conducted to statistically compare the reliability of response for different rater categories within the context of multi-rater assessments.

Evaluation of the Literature

The literature suggests that companies using multi-rater feedback systems rate their assessment systems higher than those using more traditional performance appraisal systems in areas of employee understanding, self-awareness and communication between managers and their subordinates. However, much of the data available in the literature is anecdotal or very subjective in nature. This makes it difficult to draw firm conclusions regarding how multi-rater assessment may benefit the process of conducting performance assessments in organizations.

Some studies demonstrate that managers and leaders may improve their performance (at least as it is perceived by others) after receiving feedback from multiple sources, especially if the initial feedback is negative. One explanation for this finding may be that negative feedback is more amenable to setting achievable goals than positive feedback on which there is no perceived need for action. However, several limitations are apparent in the literature

pertaining to performance improvements. First, most studies reviewed deal only with a single rater category (e.g. subordinates rating their immediate managers) which makes it difficult to generalize their results to other rater categories (e.g. colleagues, supervisors, customers). Also, no control groups were included in the research designs of the reviewed studies. This did not allow for the evaluation of situational variables as possible contributors to changes in leadership behavior. Likewise, small sample sizes may have impeded researchers in their ability to detect small behavior changes more apparent when studying larger samples. Finally, individuals participating in the studies mentioned could not be randomly assigned to experimental groups. The quasi-experimental designs that the researchers used limit the strengths of the conclusions that may be drawn pertaining to performance improvement.

Most researchers agree that, in general, self-ratings have a tendency to be inflated relative to the ratings of others. However, considerable disagreement exists as to whether this tendency impacts the reliability of self-ratings. Very little research has been conducted to confirm whether rater categories differ as to the reliability of their responses. This has interesting implications for the development of multi-rater feedback instruments. More specifically, scales reliable for colleague, subordinate, or

supervisor ratings may not be reliable for self-raters and vice versa.

The purpose of this study was to analyze the underlying constructs and psychometric properties of a multi-rater feedback instrument used for leadership development within a division of a high technology firm. Additionally, the instruments scales were tested for reliability of response overall and for each rater category.

Method

Participants

The leadership survey was administered to 750 participants of a high technology firm of approximately 14,000 employees over a three year time period. Data for this study was obtained through archival means. Demographic characteristics of the sample are not available due to the anonymous and confidential nature of the survey process. However, the firm is an engineering company specializing in the design and manufacture of semiconductor processing equipment. There were a total of 48 individuals, each of which provided a self-rating. Additionally, a total of 358 colleague raters, 285 subordinate raters, and 59 supervisor raters provided ratings for the 48 feedback participants. Each of the 48 rated individuals was rated by an average of 15 other raters.

Instrument

The instrument used as a basis for the multi-rater assessment was developed over a period of three months by organization development professionals with executive input. The organization's president was asked a series of questions over the course of three scheduled sessions to explore his expectations of leadership within his division. Once his input was collected, the list of expectations was circulated to other executives within the division. Once all input was gathered a final list was compiled and agreed to during an executive off-site meeting. What resulted from this development effort were 34 items divided into four competency groups. No assessment of the instrument's factor structure and reliability has been attempted to date. The four competency groups include: Focus on Objectives and Results (7 items), Support for Healthy and Productive Relationships (10 items), Development and Maintenance of Effective Work Processes (11 items), and Strong Sense of Business Leadership (6 items) (See Table 1 for a review of all items).

The Focus on Objectives and Results scale includes items such as "Establishes clear objectives and goals", "Holds people accountable for their actions" and "Adopts and communicates a big picture perspective".

The Support of Healthy and Productive Relationships sub-scale includes items such as "Fosters a high level of

trust among team members", "Encourages honesty and openness" and "Celebrates accomplishments". The Development and Maintenance of Effective Business Processes scale includes items such as "Gives actionable feedback", "Delegates effectively when appropriate" and "Never compromises on safety". Finally, the scale measuring Business Leadership includes items such as "Is passionate about the work and our business", "Focuses on action; makes decisions and implements quickly", and "Shows determination and perseverance when pursuing a goal".

Respondents are asked to assess the manager being rated with regard to each competency on a 5-point Likert scale of frequency where 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, and 5 = Always. Each rater category responds using a distinct form, as all items are not rated by every rater category.

Survey Administration

The administration of this survey was essentially a compilation of many individual administrations conducted over a 3 year time period. The managers being rated compiled lists of subordinates, colleagues, and supervisor(s) to whom the survey should be administered. The survey instruments were administered through the company's inter-office mail system to those identified by the manager and the manager being rated completed a self-assessment. Because the survey was anonymous, each survey

was coded to determine which rater category the respondent represented on return. Survey participants were given approximately two weeks to complete and return the survey. Once returned, numerical data (item and scale means) and rater comments were entered into a database which calculated averages and produced a report. The report was then used in a one to one feedback session with the rated manager conducted by an organization development professional.

Results

Descriptive Statistics

Item means, standard deviations, and inter-item correlations for each of the four survey sections appear in Tables 2-5. Included within these tables are scale means for all respondents. The lowest represented scale mean occurred for the Focus on Objectives and Results scale, M=3.75, which also contained the lowest item mean (item 5) for the sample, M=3.45. The highest scale mean occurred for the Sense of Business Leadership scale, M=3.96. This scale encompassed the second and third highest item means, M=4.29 and M=4.22, for items 30 and 31 respectively. The highest item mean occurred for item 18 within the Development and Maintenance of Effective Business Processes scale, M=4.33. Total scale means and standard deviations indicate a tendency for respondents to rate near the center and upper end of the rating scale.

Table 1

Item Descriptions

Item no.	Scales and Item Descriptions
	Focuses on Objectives and Results
1	Establishes clear objectives and goals
2	Routinely focuses on given objectives and goals
3	Adopts and communicates a big picture perspective
4	Has a global perspective
5	Communicates an inspiring vision
6	Establishes clear priorities
7	Holds people accountable for their actions
	Supports Healthy and Productive Relationships
8	Communicates effectively with superior when appropriate
9	Communicates effectively with direct reports
10	Communicates effectively with me
11	Listens to me when I'm talking to him/her; gives me full attention
12	Fosters a high level of trust among team members
13	Treats others with respect and dignity
14	Encourages honesty and openness
15	Expects top performance; does not tolerate incompetence
16	Celebrates accomplishments
17	Seeks feedback; wants to improve
	Development and Maintenance of Effective Business Processes
18	Does not compromise on safety
19	Makes difficult decisions
20	Makes decisions that are crisp and concise
21	Follows through on commitments
22	When appropriate, uses quality tools
23	Emphasizes and delivers quality products
24	Approaches problems with a process orientation; develops plans and outlines steps to be followed
25	Minimizes bureaucracy; works to eliminate excessive red tape and paperwork
26	Gives actionable feedback
27	Delegates effectively when appropriate
28	Encourages cross-functional interaction and teamwork
	Strong Sense of Business Leadership

29	Focuses on action; makes decisions and implements quickly
30	Is passionate about the work and our business
31	Shows determination and perseverance when pursuing a goal
32	Gets good things to happen to him/her
33	Understands and communicates the relationship between my function and the company's success in the marketplace
34	Exemplifies and reinforces the company's values

Table 2

Mean Responses and Inter-item Correlations for Items in Focus on Objectives and Results Scale

Item no.	N	M	SD	1	2	3	4	5	6
1	744	3.81	.71	--					
2	745	3.86	.76	.62*	--				
3	746	3.78	.87	.36*	.33*	--			
4	745	3.88	.87	.22*	.22*	.56*	--		
5	743	3.45	.92	.41*	.33*	.58*	.43*	--	
6	742	3.68	.78	.57*	.53*	.36*	.25*	.43*	--
7	737	3.81	.82	.35*	.36*	.21*	.21*	.29*	.40*
Total	750	3.75	.82						

Note: Values presented represent the mean response for each item.

* $p<.01$

Table 3

Mean Responses and Inter-item Correlations for Items in Support of Healthy and Productive Relationships Scale.

Item no.	N	M	SD	8	9	10	11	12	13	14	15	16
8	662	3.93	.67	--								
9	380	3.74	.77	.36*	--							
10	692	3.92	.84	.43*	.64*	--						
11	744	4.12	.84	.28*	.42*	.58*	--					
12	332	3.95	.95	.29*	.55*	.60*	.47*	--				
13	744	4.22	.81	.26*	.42*	.46*	.53*	.62*	--			
14	739	4.22	.82	.30*	.44*	.50*	.45*	.69*	.67*	--		
15	734	3.92	.79	.22*	.18*	.17*	.06	.30*	.05	.21*	--	
16	722	3.59	.88	.27*	.42*	.32*	.31*	.47*	.36*	.35*	.19*	--
17	736	3.62	.94	.23*	.46*	.47*	.38*	.54*	.44*	.49*	.23*	.47*
Total	750	3.92	.83									

Note: Values presented represent the mean response for each item.

* p<.01

Table 4

Mean Responses and Inter-item Correlations for Items in Develop and Maintain Effective Work Processes Scale

Item no.	N	M	SD	18	19	20	21	22	23	24	25	26	27
18	710	4.33	.93	--									
19	735	3.83	.75	.35*	--								
20	737	3.68	.75	.27*	.59*	--							
21	732	3.96	.75	.23*	.30*	.41*	--						
22	701	3.57	.82	.22*	.25*	.29*	.35*	--					
23	731	3.96	.74	.29*	.34*	.39*	.45*	.52*	--				
24	734	3.79	.91	.25*	.23*	.36*	.31*	.47*	.43*	--			
25	736	3.85	.80	.16*	.37*	.39*	.29*	.27*	.36*	.24*	--		
26	725	3.64	.75	.20*	.36*	.43*	.37*	.34*	.31*	.38*	.40*	--	
27	733	3.74	.79	.17*	.32*	.36*	.33*	.29*	.29*	.30*	.25*	.46*	--
28	742	3.89	.88	.26*	.27*	.32*	.34*	.38*	.36*	.45*	.28*	.40*	.47*
Total	750	3.84	.81										

Note: Values presented represent the mean response for each item.

* p<.01

Table 5

Mean Responses and Inter-item Correlations for Items in Strong Sense of Business Leadership Scale

Item no.	N	M	SD	29	30	31	32	33
29	734	3.90	.74	--				
30	737	4.29	.76	.43*	--			
31	730	4.23	.69	.49*	.56*	--		
32	705	3.71	.72	.43*	.34*	.46*	--	
33	327	3.68	.95	.42*	.43*	.37*	.42*	--
34	730	3.91	.80	.31*	.36*	.36*	.42*	.51*
Total	750	3.96	.78					

Note: Values presented represent the mean response for each item.

* $p<.01$

A review of inter-item correlations reveals high correlations overall with only several exceptions. Most notably, item 18 (Does not compromise on safety) within the Development and Maintenance of Effective Business Processes scale correlated poorly with other items in the scale. This is most probably due to the fact that item 18 is the only item within the scale dealing with safety. In addition, item 15 within the Support for Healthy and Productive Relationships scale demonstrated low inter-item correlations with other scaled items. Item 15 (Expects top performance; does not tolerate incompetence) deals directly with performance expectations, while other scaled items tend to measure aspects of the leader's effectiveness in terms of communication and how well he or she relates to other people. Despite these exceptions, the high number of significant positive correlations demonstrates that scaled items tend to measure related, if not the same, constructs.

Principal Components Analysis and Reliability Assessment

Each of the four core sections of the survey were subjected to a principal components analysis to identify the underlying constructs or factors being measured by the instrument. Principal components analysis was used in favor of factor analysis due to the methodology employed in the development of the scale. Had exploratory methods been used to develop the scales, more theoretical commitments could be made and factor analytic methodologies employed. For the

purposes of this study, each factor was assumed to vary independent of all others. Therefore, varimax rotation was used to assure the greatest chance for variables to load highly on only one factor, allowing for the clearest possible interpretations. Eigenvalues equal to or greater than 1.0 were considered meaningful for the purposes of this study.

Cronbach's alpha was then computed to test the internal consistency of each identified factor. Alpha scores are an indication of the proportion of variance in the scale scores that may be attributable to the true score (DeVellis, 1991). This analysis helps to identify those items that have a tendency to weaken a scale's internal consistency and is therefore valuable for making recommendations for revisions to the current instrument. Alpha levels of .70 were considered reliable. Reliability coefficients were also computed for each scale by rater category (self, subordinate, colleague, and supervisor). This was done to determine if the instrument's reliability varied as a function of rater group.

Overall factor structure and reliability. A total of seven factors were identified as meaningful between the four scale sections of the survey. See Tables 6-9 for a presentation of each section's factor structure and factor loadings for the items. In addition to factor loadings,

Table 6

Factor Loadings for Items in Focus on Objective and Results Scale

Item	Factor 1	Factor 2	h^2
Objective Setting, Prioritization, and Accomplishment			
q1 Establishes clear objectives and goals	.81	.19	.69
q2 Routinely focuses on given objectives and goals	.81	.15	.67
q6 Establishes clear priorities	.77	.24	.65
q7 Holds people accountable for their actions	.62	.13	.40
Maintenance and Communication of a Global Perspective and Vision			
q3 Adopts and communicates a big picture perspective	.23	.83	.76
q4 Has a global perspective	.06	.84	.70
q5 Communicates an inspiring vision	.36	.71	.65

Note: Values presented represent the factor loadings for each item.

Table 7

Factor Loadings for Items in Support for Healthy and Productive Relationships Scale

Item	Factor 1	Factor 2	h^2
Communication, Respect for the Individual, and Teaming			
q9 Communicates effectively with his/her peers	.64	.36	.54
q10 Communicates effectively with me	.73	.29	.61
q11 Listens to me when I'm talking to him/her	.75	.00	.57
q12 Fosters a high level of trust among team members	.76	.36	.70
q13 Treats others with respect and dignity	.83	.00	.69
q14 Encourages honesty and openness	.76	.22	.62
q17 Seeks feedback; wants to improve	.59	.39	.50
Upward Communication and Performance Management			
q8 Communicates effectively with his/her superior	.31	.51	.36
q15 Expects top performance; does not tolerate incompetence	-.07	.87	.76

Note: Values presented represent the factor loadings for each item.

Table 8

Factor Loadings for Items in Development and Maintenance of Effective Work Processes Scale

Item	Factor 1	Factor 2	h^2
Commitment to Quality, Follow-Through, and Effective Work Processes			
q21 Follows through on commitments	.49	.41	.40
q22 When appropriate, uses quality tools	.77	.11	.61
q23 Emphasizes and delivers quality products	.65	.32	.52
q24 Approaches problems with a process orientation	.75	.14	.59
q27 Delegates effectively when appropriate	.46	.41	.38
q28 Encourages cross-functional interaction and teamwork	.67	.25	.51
Safety, Decision Making, and Continuous Improvement			
q18 Never compromises on safety	.23	.42	.23
q19 Makes difficult decisions	.07	.84	.70
q20 Makes decisions that are crisp and concise	.23	.78	.66
q25 Minimizes bureaucracy	.21	.61	.41
q26 Gives actionable feedback	.43	.53	.47

Note: Values presented represent the factor loadings for each item.

Table 9

Factor Loadings for Items in Strong Sense of Business Leadership Scale

Item	Factor 1	h^2
q29 Focuses on action	.72	.52
q30 Is passionate about the work and our business	.73	.53
q31 Shows determination and perseverance when pursuing a goal	.76	.58
q32 Gets good things to happen to him/her	.71	.50
q33 Understand relationship between function and market success	.73	.53
q34 Exemplifies and reinforces the organization's values	.67	.46

Note: Values presented represent the factor loadings for each item.

communality figures (h^2) are also included in Tables 6-9. These figures estimate the amount of variance within any one item that may be accounted for by all of the extracted factors taken together. The highest communality figure occurred within the Support of Healthy and Productive Relationships scale for item 15 (h^2=.76). This fairly large communality demonstrates that 76% of the variance within item 15 can be accounted for by the identified factors. On the other hand, the lowest communality for the instrument was found within the Development and Maintenance of Effective Business Processes scale for item 18 (h^2=.23). This indicates that only 23% of the variance within item 18 can be accounted for by the extracted factors.

Focus on objectives and results scale. Two factors were extracted for this section and were able to account for a total of 64.6% of the scale's total variance. The first factor accounted for 47.5% of the variance and the second was able to account for the remaining 17.1%. Examination of the items loading high on each factor allowed for labeling of the factors as underlying constructs of the scale. The first factor was labeled Objective Setting, Prioritization, and Accomplishment as those items loading high measure a leader's tendency to set, prioritize, and measure the achievement of goals and objectives. The second factor was labeled Maintenance and Communication of a Global Perspective and Vision as the items making up the scale focus on establishing, communicating, and

managing to a high level vision within the context of a global organization.

Refer to Table 10 for a review of alpha statistics for the total sample and for rater categories. Alpha criterion for these factors demonstrates that the Objective Setting, Prioritization and Accomplishment scale to be reliable (α=.77). The Maintenance and Communication of a Global Perspective and Vision scale is also reliable according to the criterion (α=.77). A review of inter-item correlations and item variability for both scales demonstrated high positive inter-item correlations and good variability of response, which contributed substantially to the high alpha coefficients.

Support of healthy and productive relationships scale. Two factors were extracted from the Relationships section of the survey instrument and accounted for 57.7% of total variance. The first factor accounted for 46.9% of the total and was label as Communication and Teamwork. The items that loaded high on this factor focus on individual and group communication to colleagues and subordinates, respect for individual and cultural differences and the tendency for a leader to promote teamwork and cross-functional collaboration. The second factor was labeled Upward Communication and Performance Management and accounted for the remaining 10.8% of the variance. This factor was made up of items dealing with

Table 10

Subscale Reliability by Rater Category

Subscale	Total	Self	Sub	Coll	Sup
Focus on Objectives and Results	.81*	.65	.84*	.78*	.87*
Objective Setting, Prioritization, and Accomplishment	.77*	.68	.78*	.76*	.83*
Maintenance and Communication of Global Perspective and Vision	.77*	.57	.80*	.72*	.84*
Support for Healthy and Productive Relationships	.87*	.73*	.87*	.81*	.74*
Communication, Respect for the Individual and Teaming	.90*	.72*	.90*	.81*	.75*
Upward Communication and Performance Management	.36	.05	.39	.42	.14
Development and Maintenance of Effective Business Processes	.84*	.76*	.86*	.83*	.85*
Commitment to Quality, Follow-through and Effective Work Processes	.79*	.61	.79*	.79*	.82*
Continuous Improvement and Decision Making	.72*	.65	.75*	.70*	.67
Strong Sense of Business Leadership	.82*	.70*	.83*	.75*	.90*

Note: The values represent Cronbach alpha coefficients for each rater catgory.

*$\alpha \geq .70$

leader communication with his or her supervisor and with management of subordinate performance. Alpha results indicate that the Communication and Teamwork scale (α=.90) is highly reliable, again maintaining high inter-item correlations and considerable variability of response. The Upward Communication and Performance Management scale, however, was not found to be reliable. (α=.36). This scale's low reliability appears to be the result of being made up of only two items which are not highly correlated.

Development and maintenance of effective work processes. This section of the survey was also found to have two underlying factors, both of which were able to account for 49.9% of the variance in this section. The first factor accounted for 40.0% of the variance and was labeled Commitment to Quality and Effective Work Processes. Items loading on this scale focused on a leader's ability to make quality a high concern, meet commitments, and develop efficient methods for getting work done. The second factor accounted for the remaining 9.9% of variance and was named Decision Making and Continuous Improvement. Items loading high on this factor primarily measure the ability to make quick, effective decisions, and to continuously improve the quality of products and work. Alpha results determined that both factors, Commitment to Quality and Effective Work Processes (α=.79), and Decision Making and Continuous

Improvement (α=.72) are reliable scales. Again, good variability among responses and high inter-item correlations contributed to high alphas.

Strong sense of business leadership scale. Only one factor was extracted from the last section of the survey accounting for 51.8% of the variance in this section. Because only one factor could be extracted no label was assigned as all items seem to be representing the scale as already constituted.

Alpha statistics reveal that this single factor solution for Strong Sense of Business Leadership (α=.82) meets the criterion for reliability. Strong inter-item correlations and good variability of response again contributed to a strong alpha coefficient.

Rater Category Rating Agreement

Scale means and standard deviations for rater category are shown in Table 11. No substantial differences in response between rater categories occurred, indicating a general tendency for raters from different categories to rate individuals similarly. The least agreement occurred for the Maintenance and Communication of Global Perspective and Vision scale, although the largest rating difference was only .23 (between subordinates and supervisors). The Commitment to Quality, Follow-through, and Effective Work

Table 11

Subscale Means and Standard Deviations by Rater Category

Subscale		Self n=48	Sub n=277	Coll n=352	Sup n=59
Objective Setting, Prioritization, and Accomplishment	M	3.73	3.77	3.80	3.97
	SD	(.50)	(.65)	(.56)	(.57)
Maintenance and Communication of Global Perspective and Vision	M	3.74	3.81	3.64	3.58
	SD	(.49)	(.82)	(.67)	(.77)
Communication, Respect for the Individual and Teaming	M	3.85	3.91	3.94	4.05
	SD	(.42)	(.70)	(.58)	(.45)
Upward Communication and Performance Management	M	3.95	4.00	3.88	3.93
	SD	(.49)	(.58)	(.58)	(.59)
Commitment to Quality, Follow-through and Effective Work Processes	M	3.77	3.87	3.78	3.85
	SD	(.42)	(.60)	(.57)	(.56)
Continuous Improvement and Decision Making	M	3.99	3.92	3.79	3.92
	SD	(.45)	(.61)	(.50)	(.52)
Strong Sense of Business Leadership	M	4.03	3.99	3.98	4.08
	SD	(.43)	(.62)	(.49)	(.51)

Note: Values represent subscale means and standard deviations for individual rater categories.

Processes scale, and the Strong Sense of Business Leadership scale both demonstrated the highest levels of agreement between rater categories, neither differing more than .10. In addition, a close review of self-ratings finds no inflation, when compared to ratings of other groups. Self-ratings, however, demonstrate less variability of response overall, which may contribute to low reliability coefficients.

Scale Reliability by Rater-Category

Refer to Table 10 for a listing of alpha coefficients for each subscale by rater category (self, subordinate, colleague, and supervisor). Note that the four original scales are reliable overall. Also, the four scales are reliable across rater categories, with the exception of the Focus on Objectives and Results scale which is not reliable for self-raters. Low correlations and poor variability of response seem to be the cause for the low alpha.

Six of the seven identified subscales are reliable for all rater categories taken together. The Upward Communication and Performance Management Scale is not reliable, and this most likely due to the few items making up the subscale.

For colleague and subordinate ratings, six of the seven subscales are reliable. Again, the Upward Communication and Performance Management scale is not reliable. For supervisor ratings, five of the seven subscales are

reliable. In addition to the Upward Communication and Performance Management scale, the Decision Making and Continuous improvement scale was found to be unreliable. Review of item-total statistics for supervisors reveals that item 18 (Does not compromise safety) correlates poorly with other items in the scale, and that its removal would increase alpha to within acceptable limits of reliability (.78). Review of item-total statistics for all respondents demonstrates that removal of item 18 would raise alpha from current levels also (.74).

Only two of seven subscales are reliable for self-ratings. However, the remainding subscales did not meet the criterion for reliability. A small sample size (n=48) may account for the low alphas, however, variability of response is also low compared to the other rater categories which may have an impact on scale reliability.

Discussion

The purpose of this study was to analyze the underlying constructs and psychometric properties of a multi-rater feedback instrument used for leadership development within a division of a high technology firm.

The question of identifying the underlying constructs being measured by the instrument was answered through a principal components analysis. The four original groupings of items proved to be composed of several underlying factors which help to better explain how leadership is being

measured by the current instrument. Results indicated that the Focus on Objectives and Results scale was comprised of two factors. The first involves the establishment of clear objectives and goals, the prioritization of those goals, and holding others accountable for meeting those goals and objectives. The second involves having a "big picture" perspective, taking into account the global nature of business, and establishing a clear long term vision.

The Support of Healthy and Productive Relationships scale was found to consist of two primary factors. First, leaders scoring high in this area focus on effective communication (i.e. verbal one on one, attentive listening, encouragement of honesty and openness) and respect for others (i.e. rewarding accomplishments, treating others with dignity). Second, effective leaders in this area are able to openly and effectively communicate with their superiors and manage work in such a way that top performance is expected.

The Development and Maintenance of Effective Business Processes scale was also found to consist of two underlying factors. First, leaders that score high in this area emphasize quality of products and procedures, planning, collaboration and the wise use of resources. Second, they are willing to be decisive and clear when difficult decisions must be made, and they give feedback that may be acted on.

Only one factor emerged for the final scale, Strong Sense of Business Leadership. Therefore, leaders scoring high in this area are determined and passionate about their work, understand the business and its environment, are decisive and driven, and are able to relate their work to the overall success of the organization.

The second research purpose involved testing the reliability of the underlying scales of the survey instrument for all respondents and for each rater category independently. This was done to determine if the instrument's subscales reliably measure their constructs and whether scale reliability varies according to rater category.

For all respondents taken together, six of the seven subscales identified in the principle components analysis were found to be reliable. The Upward Communication and Performance management subscale was not reliable. This result was most likely due to being composed of only two items that correlate poorly with one another. The addition of items related to communicating with one's supervisor and managing the work of subordinates should help to improve the scale's reliability. Additionally, the scale might benefit from a reevaluation of keeping items related to self-manager relations and self-subordinate relations together, as currently the correlations between the two are very low. With the exception of this scale, however, results tend to

confirm the reliability of the instrument's scales as currently constituted.

For the different rater categories, reliability of subscales did differ somewhat. Only two of the seven subscales were reliable for self-ratings, six were reliable for colleague ratings, five were reliable for supervisor ratings, and six were reliable for subordinate ratings. Again, the Upward Communication and Performance Management scale was found to be consistently unreliable across rater categories due to low correlations between only two items making up the scale. The Continuous Improvement and Decision Making scale was unreliable for supervisor ratings, due to low correlations between item 18 (Does not compromise on safety) and the other scale items. This low correlation seems to be due to the fact that item 18 is the only item in the scale related to safety. Removal of item 18 from this scale is recommended, as this would assure reliability of supervisor ratings and would increase alpha for all other rater categories.

Self-ratings were found to be less reliable across subscales than any other rater category. Low inter-item correlations and considerably lower variability of response seems to account for the low alphas. This lack of variability may be due to sample size, as only 48 self-ratings were available. However, supervisor raters were also represented in a small sample (n=59) and were found to

be more reliable than self-raters. Also, item-total statistics did not identify any items for removal that would improve alpha coefficients substantially to meet the criteria for reliability for any of the unreliable scales. Small improvements to scale reliability for self-raters might be made by item-removal, but additional item removal would lower reliability for the instrument overall. A review of scale means does not show self-ratings to be inflated relative to other raters. However, self-ratings varied less than ratings of others, and therefore tended to confirm the assertion of previous researchers that self-ratings are generally less reliable than ratings made by others. In the context of multi-rater assessment, however, this result is not necessarily negative. One of the purposes of multi-rater feedback is to allow for a comparison of self-perception with the perception of others. As noted in the literature, when discrepancies in self-other perceptions were discovered, changes were made to bring self-perception more into line with those of others. This comparison, therefore, aids participants in discovering where improvements need to be made. When no discrepancies were noted, however, performance did not change. Therefore, self-perceptions may be more important for their comparative value than for rating reliability.

Implications

These results have important implications for the current instrument and for multi-rater assessment overall. First, a study of the underlying constructs of the instrument's scales has provided a better definition of what leadership behaviors are being measured. Second, the four scales of this instrument and their underlying subscales seem to be reliably measuring leadership as defined by their constructs. This allows for more confidence in the interpretation of results than was possible before. Revisions to (or the elimination of) the Upward Communication and Performance Management subscale is recommended. In addition, all of these scales are reliable for supervisors, colleagues, and subordinate raters.

Finally, self-ratings were shown to be less reliable than other ratings, demonstrating less variability of response overall. This has implications for multi-rater assessment in that an instrument administered to self and others will likely be more reliable for others than for self-ratings. This likelihood should influence researchers and practitioners in how results are interpreted and aid them in making comparisons based on self-other rating discrepancies.

Future Research and Conclusion

Due to the anonymous nature of the assessment no demographics other than rater category were collected. This

eliminates the possibility of examining potential rating distinctions based on gender, experience level, management level, and numerous other categories. Also, the numbers of self ratings and supervisor ratings represented in the sample were considerably less than subordinate ratings and colleague ratings. This may have and impact on variability of response and the reliability classifications of scales and subscales. Future studies should work to increase the number of self and boss scores included.

In summary, the analyses conducted as part of this study, helped to identify those underlying constructs being measured by a multi-dimensional survey designed to assess leadership performance. Also, an assessment of scale reliability found the instrument to be reliable with minor revisions. Finally, an assessment of reliability ratings by rater category found all to be reliable except self-ratings which tended to demonstrate poor variability of response. It is hoped that this study of a multi-rater assessment instrument will aid multi-rater feedback administrators to improve their measurement methodologies, providing more meaningful feedback to leaders in their organizations, and improving leadership overall.

References

Aamodt, M.G. (1991). Applied Industrial/Organizational Psychology. Pacific Grove, CA: Brooks/Cole Publishing.

Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self and follower ratings of leadership. Personnel Psychology, 48, 35-59.

Bohl, D.L. (1996). Minisurvey: 360-degree appraisals yield superior results, survey shows. Compensation & Benefits Review, September, 16-19.

DeVellis, R.F. (1991). Scale development: Theory and applications. Newbury Park, CA: Sage.

Edwards, M.R., & Ewen, A.J. (1996). 360 feedback: The powerful new model for employee assessment and performance improvement. New York, NY: Amacom.

Lepsinger, R., & Lucia, A.D. (1997). The art and science of 360 feedback. San Francisco, CA: Pfeiffer.

London, M., & Smither, J.W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance related outcomes? Personnel Psychology, 48, 803-839.

Murphy, K.R., & Davidshofer, C.O. (1997). Psychological testing: Principles and applications. Englewood Cliffs, NJ: Prentice Hall.

Nilsen, D., & Campbell, D.P. (1993). Self-observer rating discrepancies: Once an overrater, always an overrater? Human Resource Management, 32, 265-281.

Reilly, R.R., Smither, J.W., & Vasilopoulos, N.L. (1996). A longitudinal study of upward feedback. Personnel Psychology, 49, 599-612.

Smither, J.W., London, M., Reilly, R.R., & Millsap, R.E. (1995). An examination of the effects of an upward feedback program over time. Personnel Psychology, 48, 1-34.

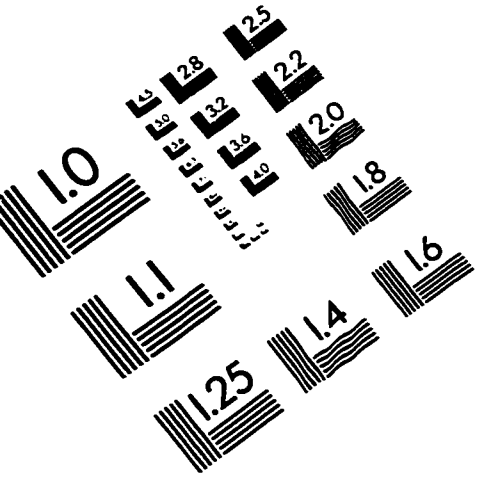
Thomas, S.L., & Bretz, R.D. (1994). Research and practice in performance appraisal: Evaluating employee performance in America's largest companies. SAM Advanced Management Journal, 4, 28-34.

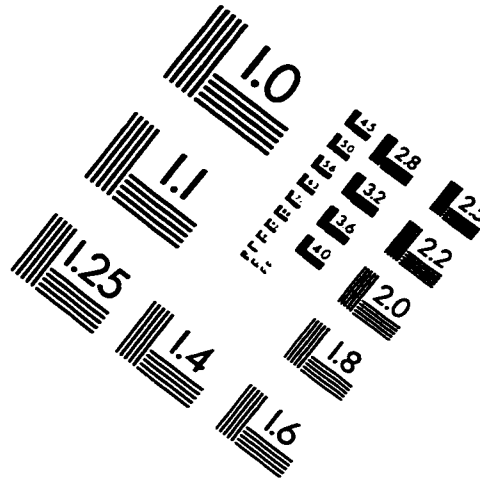
Wanguri, D.M. (1995). A review, an integration, and a critique of cross-disciplinary research on performance appraisals, evaluations, and feedback: 1980-1990. Journal of Business Communication, 32, 267-293.

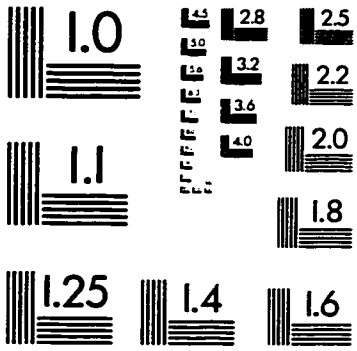
Wohlers, A.J., Hall, M-J., & London, M. (1993). Subordinates rating managers: organizational and demographic correlates of self/subordinate agreement. Journal of Occupational and Organizational Psychology, 66, 3-11.

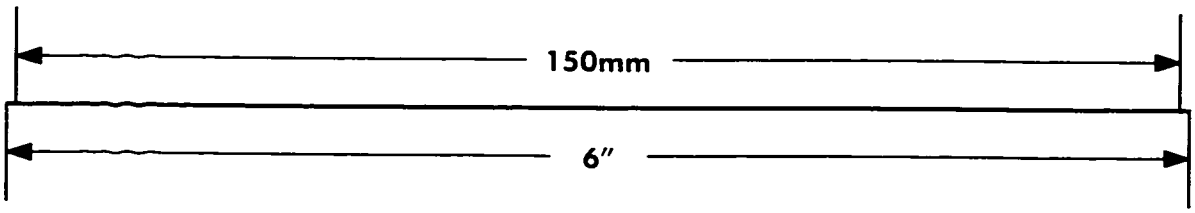
Yammarino, F.J., & Atwater, L.E. (1993). Understanding self-perception accuracy: Implications for human resource management. Human Resources Management, 32, 231-247.

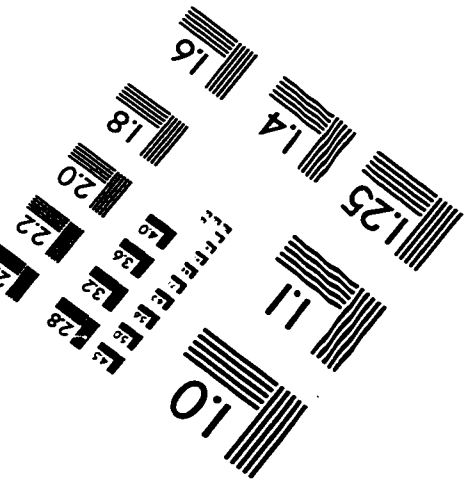
IMAGE EVALUATION
TEST TARGET (QA-3)

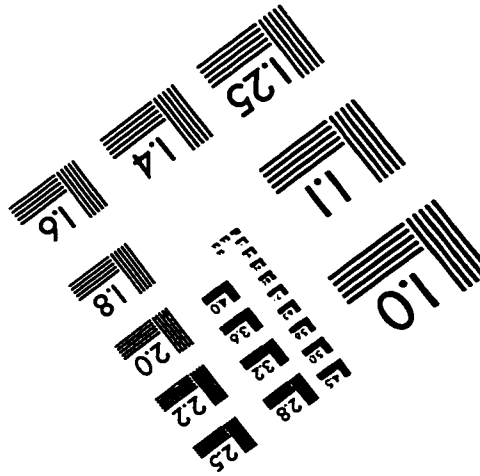












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989