A Validation Study of the Federal Aviation Administration's Assessment Center for Facility Chiefs

Summer 1981

Nancy K. Blagg

University of Central Florida

Find similar works at: https://stars.library.ucf.edu/rtd

University of Central Florida Libraries http://library.ucf.edu

Part of the Industrial and Organizational Psychology Commons

STARS Citation


https://stars.library.ucf.edu/rtd/538

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
A VALIDATION STUDY OF THE
FEDERAL AVIATION ADMINISTRATION'S
ASSESSMENT CENTER FOR FACILITY CHIEFS

BY

NANCY KAY BLAGG
B.A., Eckerd College, 1974

THESIS

Submitted in partial fulfillment of the requirements
for the Master of Science Degree in Industrial Psychology
in the Graduate Studies Program of the College of Social Sciences
University of Central Florida
Orlando, Florida

Summer Term
1981
Table of Contents

Introduction .............................................. 1
  Present Study .................................... 23
  Hypotheses ..................................... 24
Methods .................................................. 25
Results .................................................. 30
Discussion ............................................ 35
Appendix A ............................................. 45
Appendix B ............................................. 67
Appendix C ............................................. 72
References ............................................ 78
Acknowledgements

It is with grateful appreciation that I acknowledge the support and professional skills of my advisor, Dr. Wayne A. Burroughs and my committee members Dr. Cabot L. Jaffee and Dr. John M. McGuire. The two years I spent in the graduate program, at what was then Florida Technological University, provided me a sound data base to operate in the real world of industrial/personnel psychology. I continue to appreciate the enriching experience of the hands-on approach in both the Industrial and Community/Counseling Masters Program. My psychology professors brought to the classroom a wide range of outside experiences as counselors and consultants which further broadened each student's knowledge and understanding of how theory is applied in the public and private sector. It was this positive approach to education that enabled me to "hit the ground running" when I became the personnel psychologist at FAA headquarters in November 1976. My thanks go to the people that made my graduate studies both interesting and very enjoyable.

My thanks also go to my colleagues at FAA for their support and cooperation in this study. A great many people assisted in the data gathering, editing, and typing. My special thanks to Mr. Larry Covington, Mr. Tony Amato and Ms. Gale Tolson of the Executive Personnel Staff. Grateful appreciation is extended to numerous
managers and facility chiefs in the Air Traffic Service for their assistance in data collection. Special thanks to Mr. Marty Lynn for his expertise in SPSS data analysis and to Ms. Yvette Jefferson for her typing.

A very special thanks is extended to my husband, Dr. J. Keith Kalinowski, who provided me with love and encouragement, as well as outstanding editorial skills!

Finally, I would like to thank my parents, Alva and Esther Blagg, for their love and faith that I would "get it done, someday!"

I dedicate this work to them.
Government and private business and industry have long been faced with finding equitable and efficient methods for determining which employees should be selected or promoted to positions within the organization. Most companies or agencies have relied upon traditional methods such as interviews, paper and pencil tests, and supervisory appraisals to evaluate potential (Byham, 1970). In the last 25 years an increasing number of organizations have begun to incorporate the assessment center process as part of their selection systems. Jaffee and Cohen (1980) estimated that over 1000 commercial businesses and 50 government agencies nationwide currently use this technique for selection or developmental purposes.

Exactly what is meant by the term "assessment center"? An assessment center is a process in which individuals participate in a series of individual and group simulation exercises designed to measure skills or abilities that have been previously identified by a job task analysis as necessary for successful performance in a particular target position. The candidate's behavior is observed, recorded, classified and rated by several trained assessors. Candidates normally receive feedback concerning their strengths and weaknesses while management uses the information to make selection and development decisions.

Serious concerns have been raised by practitioners regarding the need for standards or guidelines for the users of the method. The Third International Congress on the Assessment Center Method, meeting
in Quebec in May, 1975, endorsed the first set of guidelines. Recent developments concerning federal guidelines related to testing led to revisions in the assessment center guidelines in 1980. The assessment center and its essential elements are defined by the Task Force on Assessment Center Standards (1980) as follows:

An assessment center consists of a standardized evaluation of behavior based on multiple inputs. Multiple trained assessors and techniques are used. Judgments about behavior are made, in part, from specially developed assessment simulations. These judgments are pooled by the assessors at an evaluation meeting during which all relevant assessment data are reported and discussed, and all the assessors agree on the evaluation of the dimensions and any overall evaluation that is made.

The following are the essential elements which are necessary for a process to be considered an assessment center:

1. Multiple assessment techniques must be used. At least one of these techniques must be a simulation. A simulation is an exercise or technique designed to elicit behaviors related to dimensions of performance on the job requiring the participants to respond behaviorally to situational stimuli. The stimuli present in a simulation parallel or resemble stimuli in the work situation. Examples of simulations include group exercises, Inbasket exercises, interview simulations, Fact Finding exercises, etc.
2. Multiple assessors must be used. These assessors must receive training prior to participating in a center.

3. Judgments resulting in an outcome (i.e., recommendation for promotion, specific training or development) must be based on pooling information from assessors and techniques.

4. An overall evaluation of behavior must be made by the assessors at a separate time from observation of behavior during the exercises.

5. Simulation exercises are used. These exercises are developed to tap a variety of predetermined behaviors and have been pretested prior to use to insure that the techniques provide reliable, objective and relevant behavioral information for the organization in question. The simulations must be job related.

6. The dimensions, attributes, characteristics, qualities, skills, abilities, or knowledge evaluated by the assessment center are determined by an analysis of relevant job behaviors.

7. The techniques used in the assessment center are designed to provide information which is used in evaluating the dimensions, attributes or qualities previously determined.

The Task Force on Assessment Center Standards (1980) goes on to say that the following kinds of activities do not constitute an assessment center.

1. Panel interviews or a series of sequential interviews as the sole technique.
2. Reliance on a specific technique (regardless of whether a simulation or not) as the sole basis for evaluation.

3. Using only a test battery composed of a number of pencil and paper measures, regardless of whether the judgments are made by a statistical or judgmental pooling of scores.

4. Single assessor assessment (often referred to as individual assessment) - measurement by one individual using a variety of techniques such as pencil and paper tests, interviews, personality measures or simulations.

5. The use of several simulations with more than one assessor where there is no pooling of data; i.e., each assessor prepares a report on performance in an exercise, and the individual reports (unintegrated) are used as the final product of the center.

6. A physical location labeled as an "assessment center" which does not conform to the requirements noted above.

The research and evaluation of the assessment center method has been extensive. Jaffee and Cohen (1980) estimated that over fifty definitive research studies, over fifty descriptions and commentaries, ten research or review articles and reports, and seven books have been published specifically about the assessment center process. It is difficult to estimate the number of internal organizational reports and unpublished studies that have been produced. Jaffee and Cohen (1980) place the number of assessment center treatments that have been written in the last 15 years at more than 150.
Support for the reliability and validity of assessment centers is substantial and comes from many different sources. However, many of the assessment center's strongest supporters are also its greatest critics. This pragmatism issues from a strong motivation in the research and user community to investigate and improve a process that has revolutionized the approach to evaluation and selection.

A review of the criterion-oriented predictive validity studies of the assessment center process will provide the research base for this paper. MacKinnon (1975) lists the different criteria of success that have been employed in predictive validity studies of the assessment center method. The following criteria have been measured against the component elements and the overall ratings and predictions of the assessment center: (a) performance ratings in jobs for which the candidate was assessed; (b) job progress such as promotions, increases in salary, etc.; and (c) job potential or ratings of the likelihood of future progress in the organization.

The first industrial use of the assessment center method occurred in 1956 when American Telephone and Telegraph Company (AT&T) undertook a massive research effort to gain insight into the management development process and to identify the personal characteristics and skills necessary to an individual's success in the Bell System companies. The Management Progress Study (Bray & Grant, 1966; Bray, Campbell & Grant, 1974; Grant, Katkovsky & Bray, 1967) is the most significant validity study ever conducted on assessment centers. The criteria variables of salary and advancement were completely
uncontaminated by the predictor variable because the assessment results were not communicated to either the assessors or to management. All information was held for research purposes only.

Over a four year period the company processed 422 men from six Bell System companies through a three and one-half day assessment center. Approximately two-thirds of the sample were recruited as management trainees immediately after college; one third had been originally employed in non-management positions and had advanced into management early on in their careers. The evaluation techniques included clinical interviews, work samples, paper and pencil tests, and participation in group problems and leaderless group discussions. At the time of assessment, half (49.6%) of the college hires were judged as having potential for middle management, and over a quarter (28.4%) of the non-college men were predicted to achieve that level. The assessment center ratings and predictions were compared with the candidate's job progress eight years later. The correlation obtained for the relationship between the predictor and the level of management reached was .44 for the college men. For the college group, 48% of those who were predicted to make middle management did so. However, 11% of those who were not predicted to reach middle management did make third level management or above. The correlation obtained for the non-college group was much higher, \( r = .71 \). For this group, 32% of the non-college men who were predicted to make middle management reached the third level of management or above. Only 5% of those who were predicted to not make middle management attained that level in the organization. The correlation of staff judgment with salary
progress was .49 for the college group and .54 for the non-college group. Significant correlations ranging from .39 to .52 were obtained for the relationship between assessment center ratings and salary increments for over 200 candidates (Grant & Bray, 1966).

Expressing the predictive validity results in terms of successful determinations of potential, 31 (82%) of the 38 men in the college sample who had made middle management had been correctly identified. For the non-college men, 15 (75%) of the 20 men who made middle management were correctly identified. Moreover, 68 (94%) of the 72 men in the combined samples who did not advance beyond the first level management were correctly identified (Dunnette, 1971).

Michigan Bell Telephone Company was one of the six Bell System companies included in the Management Process Study. In 1958 Michigan Bell and AT&T modified the research assessment procedures to evaluate the management capabilities of craftsmen. The clinical, personality and projective procedures used in the original research effort were not employed. Michigan Bell's assessment center was the prototype of the assessment centers currently used in the Bell System to evaluate 10,000 people annually, as well as the model for assessment centers throughout government and business today (Huck, 1973, 1977).

A number of studies have provided information concerning the effectiveness and validity of the Bell System assessment centers. Huck (1973, 1977) reviewed a study by Michigan Bell (1962) which compared the first 40 men assessed and promoted with the last 40 men promoted before the assessment center program began. The findings showed that approximately two-thirds (62.5%) of the assessed group
were rated "better than satisfactory" in job performance, in contrast to only one-third (35%) of the group not assessed. Further 67% of the assessed group so rated were determined by management appraisal to have the abilities required for the next level of management. On the other hand, only 35% of the non-assessed group were deemed to have demonstrated this potential. These results suggest that the assessment process may better predict a candidate's potential to perform in an upper-level management position rather than in a first-line supervisory position (Huck, 1973, 1977). Other studies (Campbell & Bray, 1967; Finley, 1970; Huck, 1974; Huck & Bray, 1976; Jaffee, Bender, & Calvert, 1970) have also supported this finding.

In another AT&T study, Moses (1973) reported a correlation of .44 (P<.001) between final assessment center ratings and the criterion was obtaining two or more promotions since assessment. The study involved 5943 candidates assessed between 1960 and 1963. The criterion was satisfied by 41% of the candidates rated "more than acceptable," by 22% of those rated "acceptable," by only 12% of those rated "questionable," and by 4% of those "not acceptable."

Moses and Boehm (1975) reported good validity for both male and female candidates in an AT&T assessment center that used assessor teams composed solely of line managers. The correlation of overall assessment rating with advancement was .37 for women and .44 for men. Various other studies add further testimony to the success and effectiveness of the assessment center process in the Bell System (Bray & Campbell, 1968; Grant & Bray, 1969; Grant, Katkovsky, & Bray, 1967; Huck & Bray, 1976; Moses & Wall, 1975).
Large companies other than AT&T have also found evidence for the predictive validity of operational assessment centers. This roster includes IBM, Sears Roebuck, Standard Oil of Ohio (SOHIO), General Electric, J.C. Penney, Union Carbide and the Wickes Corporation.

Wollowick and McNamara (1969) reported a study involving 94 lower and middle level managers nominated to participate in IBM's assessment program. The candidates were designated as "having above-average potential for advancement." In spite of the restriction of range, the researchers found a correlation of .37 ($p<.001$) between the overall assessment rating and the criterion of increase in managerial responsibility three years after assessment.

Hinrichs (1969) conducted a study of the IBM program in which separate predictions were made for 47 lower level management candidates based on judgements made by an assessment center staff and judgements made by managers. Hinrichs noted that for this sample, managers' ratings using traditional methods of deriving predictions from information contained in personnel files had a correlation of .46 with the overall assessment center ratings.

One year later the assessment center ratings and the management potential ratings were correlated with the criterion of job level attained. The assessment center predictions had a correlation of .26 with the criterion and the managerial predictions had a correlation of .32, both significant at $p<.05$.

In an 8-year followup of this study (Hinrichs, 1978), the correlations were even higher. The assessment center predictions had a correlation of .46 with job level attainment and the managerial
potential ratings had a correlation of .55 with the criterion.

The separate characteristics or skills measured in the assessment center correlated moderately well with the criterion at the one-year point. Correlations ranged from .27 to .46. Greater correlations, ranging from .34 to .69 were obtained at the 8-year point. Four managerial skill components (energy level, administrative ability, written communications and planning and organizing) did not correlate significantly with the criterion when measured at either year 1 or 8. Still, this study clearly shows significant predictive validity over an 8-year period for the assessment center. The study also raised some interesting questions as to why the managerial review of the personnel files and the predictions made by this process did as well as the assessment center. Other studies (Bray & Grant, 1966) indicate that the predictive power of the assessment center may be strongest 8 or 10 years after the fact or for positions several levels above the target job for which the candidate was evaluated.

In another IBM study, the predictive validity ratings were summarized for eleven different assessment groups, appraising 479 managers. When compared with a variety of criteria such as position level, salary, change in position level, and increase in salary, the assessment center ratings showed significant correlations for eight of the eleven groups. The coefficients ranged from .29 to .63 with a median correlation of .35 (Dodd, 1971). In yet another study, Kraut and Scott (1972) reviewed the career progress of 1,086 non-management candidates who had been observed at an IBM assessment center one to six
years previously. Substantial correlations were reported between assessment ratings and two major organizational criteria, second level promotions and demotions from management.

Extensive research has also been conducted and reported on the Sears Roebuck assessment center program (Bentz, 1967, 1971, 1980). In his review of the Sears literature, MacKinnon (1975) stated:

Assessment ratings based on tests, on an Inbasket, and on group exercises have been correlated with on-the-job performance ratings made by personnel directors, by the candidate's supervisors, and by the candidate himself, as well as with such progress criteria as indices of job mobility and of salary progress. Pages of significant (concurrent) validity coefficients have been presented in one report alone (Bentz, 1971), some type of validity having been shown for all components of the Sears program. (p. 18)

The latest information provided by Bentz (1980) sheds new light on the long range validity of the Inbasket and three Leaderless Group Discussion problems used at Sears. It is important to note that the high correlations reported earlier by Bentz (1967) are the result of criteria correlations obtained concurrently with the assessment center data collected in 1967. Significant validity coefficients for the Inbasket and Leaderless Group Discussions were obtained for two objective criteria, job progress and compensation increase.

Each year, from 1967 through 1969, 100 managerial trainees were assessed using the four exercises, the Inbasket and the three Leaderless Group Discussions. Several types of criterion data were
collected for the candidates over the next 8 years, including evaluations of performance during the first year of experience, performance compensation increases, and an evaluation by their immediate supervisor in 1977. Criterion measures taken concurrently indicated reasonable validity for all four exercises, but predictive validity evidence after one year, while supportive, was not as strong. The long term predictive validity (determined 7 to 9 years after assessment) indicated significant relationships between assessment center scores and various performance criteria, but in the direction opposite than expected. For example, it was reported that a demonstration of leadership during the group discussions appeared to predict lower performance ratings in subsequent years. Although the correlations are not strong, there was evidence to support the long-term predictive validity of the Inbasket. In summary, strong patterns of concurrent criterion-related validity have been demonstrated for both types of exercises, the Inbasket and the Leaderless Group Discussion. There are however, problems associated with the long range predictive validity of the Leaderless Group Discussion exercises (Bentz, 1980).

SOHIO has conducted several major studies or their assessment center program (Hardesty & Jones, 1968; Finkle & Jones, 1970). From 1963 to 1966, 122 entry-level management candidates were assessed and followed for a 2 1/2 to 5-year period (Carleton, 1970). A later sample of 109 candidates was assessed by the same program and the candidates' progress followed for a shorter time period (Finley, 1970).

As reported by Cohen, Moses and Byham (1974), the multirater-
multimethod validation study of the SOHIO program for both samples involved various predictor and criterion measures. Criteria included averages of thirteen supervisory ratings obtained from two independent raters and a composite measure of managerial progress (i.e., salary growth and promotions adjusted for initial status). Predictors included thirteen ratings each from of three methods; namely, a projective test, an interview, and rating from the assessment center staff. The assessment staff decisions were not independent; rather they were based upon assessment reports which included paper-and-pencil tests and sociometric data.

The findings of both studies confirmed that the assessment ratings of the programs were moderately predictive of managerial performance. Carleton's study (1970) of the earlier sample found a median correlation of .25, and Finley (1970) reported a median correlation of .32 for the later group. The assessment ratings were highly valid for the prediction of potential, with a correlation of .65 found for the early sample and a correlation of .63 for the later group. Carleton (1970) also found the assessment score to be highly predictive (r = .51) of managerial progress. The correlations associated with the assessment center predictions were larger than those based on the interviews and projective tests. Carleton found that the interviews only had correlations of .05 to .33 with managerial progress. Projective tests had correlations with the criterion of .11 to .25 (Carleton, 1970) and .01 to .34 (Finley, 1970).

In a related study by Thomson (1970) significant validities (r = .64, p < .001) were found between overall ratings in the assessment
center and ratings based on interviews with the candidate's supervisors from 6 to 27 months after assessment. When the criterion was a rating of further managerial potential evaluated 9 to 29 months after assessment, the predictive validity of the overall assessment rating was $r = .63$ \((p < .001)\).

In another SOHIO study (Mitchel, 1975) data on 154 managers attending an assessment center were examined for changes in validities over time. Twenty-four predictors were correlated with the criterion of salary growth measured 1, 3, and 5 years after the assessment center. Peer and assessor ratings were found to be significantly correlated with the criterion. All of SOHIO's reported validity research indicated predictive correlations which were very significant and highly consistent.

Other companies have also researched and established the validities of their assessment center programs. In evaluating the results of an assessment center at General Electric, Meyer (1972) found that those candidates assessed high in potential performed better on the job 1 1/2 years later than did those assessed low in potential, but both groups did better than a third group that was not assessed at all.

Jaffee, Bender and Calvert (1970) conducted a validation study of the managerial assessment center at Union Carbide. Extensive interviews conducted with incumbents' supervisors and subordinates enabled the researchers to compare men promoted as a result of the assessment center evaluation with those promoted before the inception of the program. Study results indicated that individuals promoted as a result of the assessment center perform better than those selected by
traditional means.

Worbois (1975) reported the predictive validity for Detroit Edison's management assessment center for 48 supervisors. Three kinds of criterion scores were used: (a) specific behavioral items normally used to review the performance of supervisors in the company, (b) graphic scales of the twelve abilities measured in the assessment center, and (c) a general overall rating. The results of the study indicated a consistent positive relationship between the criteria and the assessment center results. Every coefficient was statistically significant at $p<.05$.

Parker (1980) reported on the American Management Association's comprehensive survey of past users of AMA's assessment centers. To evaluate the validity of the assessment center the study used data from participating organizations that had measures of both assessment center performance and job performance for their participants. Although both types of data were available for only 280 of the 3395 participants assessed by AMA up to that time, the study yielded significant results. Management abilities as measured by the AMA assessment center proved to be moderately correlated with job performance with coefficients ranging from .26 to .37. Overall management ability had a correlation of .36 with overall job performance. The validity coefficients were statistically significant ($p<.001$), but lower than validities found in previous studies of the AMA assessment center (McConnell & Parker, 1972; Worbois, 1975).

As reported by Huck (1977), Slivinski and his colleagues at the Canadian Public Service Commission have published a series of technical
reports (Slivinski & Etheir, 1973; Slivinski, McDonald, & Bourgeois, 1979) placing particular emphasis on identifying managerial job functions and participants' reactions to the program. In one of the few validation studies for a national government assessment center reported in the literature (Slivinski, Grant, Bourgeois, Pederson, & McCloskey, 1978) data were examined for two groups of candidates evaluated for first line management positions in a Canadian federal department. The relationships of the assessment dimensions were compared to singular, composite and multiple criterion measures of job performance. Four criterion measures were used as global indicators of success. Two of them were defined in terms of career or managerial progress (salary and position level) and the other two were categorized in terms of job performance (departmental appraisal and supervisors' overall performance evaluation). The criterion data were gathered at three different times (1 1/2, 3, and 4 1/2 years) after the assessment center for the first group of 45 candidates. Data were gathered at 2 and 3 1/2 years for the second group of 72 candidates.

For the initial group, higher significant relationships were found between the overall evaluation of success probability at the assessment center and the series of global measures of career progression and job performance. Correlations ranged from .42 to .53 (p<.01). The relationship between various managerial dimensions measured at the assessment center and on the job were significant, but confined to the data collected at the 3 year point. Significant correlations then ranged from .40 to .51 (p<.01). For the second group, significant relationships were found with global measures of career progression
(correlations ranged from .26 to .36) but fewer relationships were found with measures of job performance (correlations ranged from .24 to .41). Slivinski et al. (1978) echoed Moses (1971) and Huck (1973) in stating that the overall rating given at an assessment center has proven to be a robust predictive measure when correlated with global measures of job success.

Assessment center validity studies have been examined in numerous reviews (Bray, 1976; Bray & Moses, 1972; Cohen, Moses & Byham, 1974; Dunnette, 1971; Howard, 1974; Huck, 1973; Huck, 1977; Jaffee & Cohen, 1980; Kraut, 1972; MacKinnon, 1975; Norton & Edinger, 1978). The literature has generally been interpreted as being supportive of the validity and usefulness of the assessment center method.

For example, in Byham's (1970) survey of 20 companies that operated assessment centers, 22 studies showed assessment to be more effective than other approaches and one study showed it to be as effective. None of the studies he reviewed showed the assessment center to be less effective. The studies exhibited correlations ranging as high as .64 between center prediction and achievement criteria such as advancement, salary grade and performance ratings. Cohen et al. (1974) reviewed the literature focusing on the predictive accuracy of the overall assessment rating, a global variable common to all the assessment studies. Eighteen research studies, conducted from 1964 to mid 1972, consistently showed assessment performance to be related to several criteria: the predictive accuracy was highest for job potential (median $r = .63$), followed by progress (median $r = .40$), then job performance (median $r = .33$).
Although the majority of predictive validity studies report favorable results, and most reviews of studies concluded that the assessment center has value as a predictor of various job criteria, the problems associated with this type of research should be examined. Except for the previously discussed AT&T research conducted as part of the Management Progress Study (Bray & Grant, 1966), and Hinrich's (1969, 1978) studies conducted at IBM, all the operational validity studies mentioned above were subject to various methodological problems. These included statistical restrictions of range, criterion contamination, and possible problems due to the invalidity of either the predictor or criterion measurements. The most common problem facing the researchers was criterion contamination. It was difficult to measure the varying degrees of access that line management had to the assessment reports in these studies. There has always been the question of how much influence or "self-fulfilling prophecy" was operating on the criterion measure as a result of the assessment center evaluation.

Perhaps the most serious problem to consider in any validity research effort is the appropriateness or validity of the criterion itself. Klimoski and Strickland (1977) cite Wallace (1974) when they state:

While much attention has been given to predictor (assessment center) characteristics as they influence prediction hit rates, little emphasis has been placed on what may be the heart of the matter, the differential predictability and appropriateness of various criteria used in assessment center research. (p. 353)
Klimoski and Strickland (1977) reviewed published validation studies of assessment centers with respect to criteria used, staff composition and the organizations involved. They state that, "Regardless of the center format used, these results have been impressive, positive and consistent." Nonetheless, the authors also noted that of over 90 studies reviewed for their 1977 article, they could find few validity studies per se being published after 1972. They pointed out that there was also a "curious homogeneity" in the criteria used for validation research. The authors criticized the emphasis placed on selecting indices of job advancement to the exclusion of other potential criteria by stating, "Salary growth or progress, promotions above first level, management level achieved and supervisor's ratings of potential have been popular solutions to the criterion problem." (p. 354)

Klimoski and Strickland (1977) developed a table to show that out of 17 studies they reviewed, six used a criterion of management progress (management level achieved, salary growth, number of promotions and/or increase in responsibility) to show validity. These studies were Bray and Grant (1966), Hinrichs (1969), Kraut and Scott (1972), Mitchel (1975), Moses and Boehm (1975), and Wollowick and McNamara (1969). Two studies (Campbell & Bray, 1967; Carleton, 1970), used ratings and progress as criterion measures. Five of the 17 studies reviewed (Byham & Wettengel, 1974; Ginsburg & Silverman, 1972; McConnell & Parker, 1972; Thorenson & Jaffee, 1973; Worbois, 1975) used performance ratings taken concurrently with the assessment center ratings. Only four studies (Bray & Campbell, 1968; Huck & Bray, 1976;
Jaffee, Bender & Calvert, 1970; Thomson, 1970) collected criterion ratings at various points after the assessment center. Klimoski and Strickland (1977) criticized taking performance criteria concurrently with the assessment center ratings since this may subject them to criteria contamination. They also pointed out that some of the criteria used may have had less to do with managerial performance or effectiveness than they did with managerial adaptation or politics. They criticized the use of salary and advancement as criteria since these are often quite dependent on forces far removed from a candidate's abilities or effectiveness. Klimoski and Strickland cited Cohen et al. (1974) in their discussion of a study by Carleton (1970). Carleton found a correlation of .63 between assessment center prediction of potential and ratings (obtained from higher management) of potential for higher management, but a correlation of only .33 between assessment center predictions and actual job performance. It is evidence such as this that led them to conclude that "there is a great need for predictive validity studies of assessment centers that use criteria other than those of advancement; what is needed are predictive validity studies of performance."

The use of performance ratings as criteria is not without problems. MacKinnon (1975) stated:

Generally it has been assumed, because of the possibility of contamination of the criteria by predictions, that validities have been spuriously high. There are, however, also reasons to think that in some instances they may have been spuriously low
due to invalidity of the criteria. The high percentage of unqualified among candidates nominated by supervisors for assessment suggests that supervisors' judgments are far from infallible and their ratings of job performance questionable. (p. 20)

Cohen et al. (1974) reviewed an AT&T (Bray & Campbell, 1968) study which assessed sales performance rather than managerial skills. Several precautions to minimize criterion bias were taken in this study. Individuals were hired without reference to their assessment performance and the scores were used for research purposes only. A detailed job analysis provided realistic criterion measures based on actual job performance. Three sources provided criterion ratings after the subjects had been on the job 6 months. Ratings were obtained from supervisors, trainers and from a specially trained team of sales reviewers who had the opportunity to observe each subject handling several actual sales contracts.

The validity results obtained in this study indicated strong relationships between assessment ratings and job performance. The correlation between the overall assessment rating and the field review by the special raters was .51. This is a very positive result. However, the ratings made by the supervisors and the trainers were unrelated to the assessment center ratings. Cohen et al. (1974) stated that:

These findings raise some serious methodological consideration where supervisors, trainers, or others rate job success of assessed groups for the purpose of validity analyses. In
general, such ratings tend to be subject to various constant errors. (p. 21)

Cohen et al. (1974) goes on to say:

In assessment center studies supervisors untrained in assessment methodology have often been found to be biased in rating job success and in nominating men considered promotable. As a general rule, about half of all individuals nominated as candidates to attend an assessment center are seen as not having sufficient abilities to succeed in management. Since the same supervisors who nominated candidates are often called on to evaluate subsequent performance, the uncritical use of supervisory judgment as the sole basis of validity assessment performance is a questionable procedure. As noted by Bray and Moses (1972), it is somewhat absurd to validate an expensive, high-powered (in the sense of information generated) process by a cheap, often unrealistic, but easily obtainable "criterion" measure. (p. 21)

MacKinnon (1975) further supports this concern by citing a SOHIO study by Thomson (1970). The study noted that managers serving as assessors in the program showed high agreement with psychologists in rating managerial behavior (median r = .85 between raters). But when another group of supervisors was asked to rate these same persons on the same dimensions 6 to 24 months later, its ratings were less reliable, more restricted in range, and more skewed on the side of leniency than the ratings made by the manager-assessors. MacKinnon states that "such findings must call into question the
appropriateness of using supervisors' ratings as criteria of job performance." MacKinnon qualifies this by stating:

This is not to suggest that managers cannot be effective evaluators or judges but rather that they are unlikely to make finely differentiated evaluations unless they receive the same kind of experience or training that managers who serve as assessors in assessment centers receive. (p. 20)

Cohen and Jaffee (1980) note that in order to validate any predictive measure the following must be true: (a) the measures used to predict some performance must be relevant to that performance, and the inverse, (b) the performance measures themselves must be equally relevant to the predictors. The authors stated that "the relevance of certain criterion measures is frequently suspect . . . . if management potential appraisals were valid, there would be little need for assessment center appraisals." They noted that it is relatively amazing that assessment center results have been as valid as they have. The authors indicated that most assessment center predictions are underestimates of their true validity.

The Present Study

In spite of the many problems associated with the design and use of criterion measurements, the present study was conducted using job performance ratings as a criteria to validate the predictions made in an assessment center for high level managers in a technical agency. Comparisons were made between the candidates' ratings in seven skill areas measured in the assessment center and on the job. The major difference between this and most previous validity studies is that the
criterion rating form paralleled the assessment center format. Candidates received ratings for the seven skills in each of four functional job areas: administrative duties, situational analysis, individual interaction situations, and group interaction situations. These four functional areas of the position directly corresponded to the four simulation exercises in the assessment center which were designed to sample the specific job content. The four respective exercises were the Inbasket, the Individual Problem Analysis, the Employee Counseling Exercise and the Leaderless Group Discussion.

Hypotheses

In line with previously discussed validation research, this study predicted significant correlations between the subjects' assessment center performance ratings and ratings of job performance two to four years later. The specific hypotheses are listed below:

1. Significant correlations were predicted between the specific skill scores a subject received in each of the four assessment exercises and the skill ratings the subject received for job performance.

2. Significant correlations were predicted between the subjects' exercise summary scores and job performance situation summary scores.

3. Significant correlations were predicted between the concensus scores a subject received for the seven skills rated in the assessment center and those received for job performance.

4. Significant correlations were predicted between a subject's assessment center "total score" and a total job performance rating score.
Method

Subjects

The subjects were 63 Federal Aviation Administration (FAA) GS-15 Air Traffic Facility Chiefs or Deputy Chiefs who were selected for their positions after successfully completing all four steps of a multiple-hurdles selection system which included an assessment center. All but one of the subjects were male. There was no indication that significant differences existed regarding the management experience level of the candidates at the time of assessment. All candidates had been screened for minimal managerial qualifications by their respective personnel offices and regional division chiefs. The results of the management background review (described in Appendix A) which evaluated past experience, educational level, technical and managerial training, performance evaluations and awards were essentially equivalent for all candidates referred to the assessment center.

Treatment of the Independent Variable

The 63 subjects each participated in a two day managerial assessment center either in 1977, 1978, or 1979. All three assessment centers were equivalent and comparable in design and administration. A history and detailed description of the air traffic assessment center process is provided in Appendix A.

Each candidate was observed and rated while performing in four different exercises: an Inbasket, an Individual Problem Analysis, an
Employee Counseling Exercise, and a Leaderless Group Discussion. The Inbasket was a 2 hour written exercise consisting of 40 memos that required reading and action. The Individual Problem Analysis was an exercise in which the candidate was allowed 1 1/2 hours of preparation time to select and support in writing one of several alternative courses of action. At the end of this time the candidate was required to give a short oral presentation of his or her recommendation to two role-playing assessors. The Employee Counseling Exercise required the candidate to counsel and resolve differences between two role-playing assessors. The Leaderless Group Discussion was a 1 1/2 hour exercise in which each of six candidates was given a separate position to defend and promote among the other group members. The exercises are described in greater detail in Appendix A.

**Independent Variable**

*Skill ratings.* Seven managerial skills were derived from a functional job analysis. The skills are defined in Appendix A. The skills were:

- perception and analytical ability
- organizing and planning
- leadership
- decisiveness
- judgment
- interpersonal
- oral and written communication

Using normal assessment center methodology, assessors observed the candidates in the exercises and wrote exercise reports on each
candidate's performance. The assessors discussed each candidate's performance within each exercise to arrive at the seven skill scores for each exercise. The ratings were made on a seven-point scale wherein "7" indicated "Outstanding"; a "6" meant "Well above satisfactory"; a "5" meant "More than satisfactory"; a "4" meant "Satisfactory"; a "3" meant "Less than satisfactory"; a "2" meant "Well below satisfactory"; and a "1" meant "Very weak" performance. An example of the assessment center rating form is shown as Figure 1 in Appendix A.

**Concensus skill ratings.** From the ratings in individual exercises, overall "consensus skill" evaluations were made. Each consensus skill rating was the composite individual skill rating reflecting the assessors' judgments as to the candidates "overall" skill level as demonstrated across all the exercises. The seven-point scale was also used for the consensus skill ratings.

**Assessment center total score.** The seven consensus skill ratings were added to determine the candidates' total assessment center score. Since the scores were added, the final scale ranged from 7 to 49. If a candidate received a "4" or "satisfactory" on all seven skills then he or she would achieve a total assessment center score of 28. Candidates total scores ranged anywhere from a low of 18 to a high of 49. Total scores were normally distributed.

**Exercise summary scores.** For the purpose of more easily comparing patterns of performance in the assessment center, the assessors added the rating form's vertical column of scores for each exercise. This enabled the assessors to see at a glance the strength
of a candidate's performance in each of the exercises. The written "final report" of the assessment center included a comparison of performance between types of exercises as well as a discussion of the patterns of skill performance.

**Dependent Variable**

The dependent variable used in this study was the performance rating each subject received on a specially designed job performance appraisal instrument. The criterion or job performance rating matrix form was identical in design to the predictor or assessment center performance rating form. The same seven managerial skills were rated, this time by supervisors evaluating target level job performance. The same seven-point rating scale was used. Instead of being rated on four different simulation exercises, the incumbents received skill scores according to how the supervisors perceived the incumbent's performance in four different types of actual job situations. The four performance areas rated were administrative functions, situational analysis, individual interaction situations and group interaction situations. An overview of the definition, conduct and "rules" of the performance rating process is given in Appendix B.
Statistical Analysis

A total of 78 units of data were collected for each of the 63 subjects:

\[ 2 \times 27 = 54 \]  Separate scores across four assessment center exercises and four job situations.

\[ 2 \times 7 = 14 \]  Concensus skill scores on the assessment and the performance appraisal instrument.

\[ 2 \times 4 = 8 \]  Assessment center exercise and job situation columns, summed to provide a measure of strength of performance.

\[ 2 \times 1 = 2 \]  Total scores in the assessment center and the performance appraisal instrument.

78  Total units of data to compare for each subject

A Pearson-r correlation showed the degree of correlation between predictor scores and performance scores. Regression analyses described the type of relationship that existed between significant correlations. The statistical analysis was computed using the PEARSON CORR and REGRESSION programs from chapters 18 and 20 of SPSS: Statistical Package for the Social Sciences (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

The reader should note that the skill of leadership was not measured in the Individual Problem Analysis Exercise in the assessment center, and therefore only 54 and not 56 skill scores were obtained.
Results

Table 1 shows the correlation matrix for each of the 27 skill ratings, the five consensus skill ratings, the four exercise summary ratings, and the overall rating. To simplify reporting of the results, the four predictor/criterion categories will be referred to as "administrative," "analysis," "individual," and "group," as they appear left to right in Table 1. Several significant results related to the four hypotheses emerge from Table 1.

Hypothesis I

Only four of the 27 skill rating correlations showed significance at p<.05 or better. Predictor and criterion ratings were significantly correlated with p<.05 for the following skills: perception and analytical ability as rated in group interaction situations (r = .30), and communication as rated in administrative situations (r = .28) and in analysis situations (r = .30). The skill of organizing and planning as rated in administrative functions was correlated at r = .43 with p<.01. It bears mentioning that four other skill ratings were significant at p<.10, indicating a weaker, but identifiable relationship between the ratings on the predictor and criterion variables. These skills were perception and analytical ability as measured in administrative situations, r = .22; decisiveness in analysis situations, r = .21; interpersonal as measured in group situations, r = .23; and communication as also measured in group situations, r = .23.
Table 1
Correlations Between Scores on Seven Skills Measured by Assessment Center Exercises and Job Performance Ratings

<table>
<thead>
<tr>
<th>Skills</th>
<th>Inbasket Correlated with Administrative Functions</th>
<th>Problem Analysis Correlated with Situational Analysis</th>
<th>Employee Counseling Correlated with Individual Interaction</th>
<th>Leaderless Grp. Disc. Correlated with Group Interaction</th>
<th>Concensus Skill Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptive/Analytical</td>
<td>.22</td>
<td>.12</td>
<td>.02</td>
<td>*.30</td>
<td>*.27</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>**.43</td>
<td>.06</td>
<td>.003</td>
<td>.18</td>
<td>**.36</td>
</tr>
<tr>
<td>Leadership</td>
<td>.14</td>
<td>NA</td>
<td>-.05</td>
<td>.14</td>
<td>.22</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>.17</td>
<td>.21</td>
<td>-.07</td>
<td>.18</td>
<td>*.25</td>
</tr>
<tr>
<td>Judgment</td>
<td>.14</td>
<td>.04</td>
<td>.04</td>
<td>.11</td>
<td>.08</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>.15</td>
<td>-.05</td>
<td>.15</td>
<td>.23</td>
<td>**.33</td>
</tr>
<tr>
<td>Communication</td>
<td>* .28</td>
<td>* .30</td>
<td>.05</td>
<td>.23</td>
<td>**.36</td>
</tr>
<tr>
<td>Overall Exercise Correlation</td>
<td>*.29</td>
<td>.12</td>
<td>.08</td>
<td>*.31</td>
<td>**.33 Overall Correlation</td>
</tr>
</tbody>
</table>

*p < .05  **p < .01
Hypothesis II

The relationship between assessment center ratings and performance ratings obtained in actual job situations were compared in four categories. Two categories yielded significant correlations at \( p < .05 \). A correlation of \( r = .29 \) was found in administrative situations and a slightly higher correlation, \( r = .31 \), was found in the group interaction situations. Two of the categories did not show significant correlations between assessment ratings and performance ratings. The situational analysis category had a very low, insignificant correlation of \( r = .12 \). The individual interaction category correlation, relating scores between a subject's Employee Counseling Exercise scores and those received on the job for individual interactions, was still lower, with \( r = .08 \).

Hypothesis III

Significant correlations were found for five of the consensus skill scores. Correlations for the skills of organizing and planning \( (r = .36) \), interpersonal \( (r = .33) \), and communication \( (r = .36) \) were significant at \( p < .01 \). Correlations significant at \( p < .05 \) were found for the skills of perception and analytical ability \( (r = .27) \), and decisiveness \( (r = .25) \). The correlation for a sixth skill, leadership, approached \( p < .05 \) with a correlation of \( r = .22 \). The relationship between predictor and criterion ratings was very low for the skill of judgment \( (r = .08) \).
Hypothesis IV

A moderate correlation of $r = .33$, significant at $p < .01$ was found relative to the subjects' assessment center overall score and total performance rating score.

Regression Analysis

Regression equations were computed for all significant correlations. A great deal of confidence cannot be put into the regression equations since the standard estimates of error were high, ranging in value from .68 to .96 for coefficients based on the seven-point rating scale. The standard estimates of error for the two significant exercise summary score correlations were 4.6 and 5.1 on the 7 to 49 point rating scale. The standard estimate of error for the overall correlation was 4.4 on the same scale. The prediction equations all indicated an interesting pattern of relationship between the predictor and criterion variables.

In general, for any given skill score within an exercise, or for any given skill consensus score, it appeared as though candidates scoring in the "satisfactory" range in the assessment center were actually performing a little better than that on the job, i.e., closer to "more than satisfactory." On the other hand, candidates scoring in the "much more than satisfactory" range were rated closer to "more than satisfactory" for actual job performance. Candidates scoring "more than satisfactory" in the assessment center tended to be scored "more than satisfactory" for job performance also.

Before the reader concludes that a "satisfactory" score in the assessment center underpredicts actual job performance, and a
"much more than satisfactory" score overpredicts actual job performance, a word of caution is advised. These relations may appear for a number of reasons. The reader should not be led to conclude that the predictor instrument necessarily has a "weakness." The observed relationship may be due to problems with the criterion instrument or to the restriction of range in performance scores. The regression equation was computed to describe or better understand the relationship between the assessment center ratings and the job performance ratings. The equation may describe the line that "best fits" the distribution of scores, but its slope may be altered dramatically by just a few scores or a seemingly minor trend in scores. The various possibilities are considered further in the "Discussion."

Tables 2, 3, 4, 5 and 6 report the means, standard deviations, and standard errors of the means for the assessment center ratings and the job performance ratings in each of the four exercise/job situations and in the consensus skill rating column. The tables are found in Appendix C.
Discussion

Skill Ratings

On the whole there was more variability found in the subjects' assessment center skill ratings across the four exercises than was displayed in their performance ratings. With the exception of the Employee Counseling Exercise/Individual Interaction column, the means of all the performance rating measurements were a half a point higher than the means of the assessment scores. The standard deviations of the performance ratings were smaller, indicating a greater restriction of range. This is understandable and can be attributed to several factors. In the assessment center each of the four ratings given for a skill was based on a one-time observation of a candidate's performance in a single exercise. For the most part, candidates were consistent in their level of skill demonstration. Occasionally, however, a candidate's score in a particular exercise may have differed greatly from the scores received in the other three exercises. For example, low scores may have come during the candidate's first assessment center exercise when he or she was not "warmed up," or the candidate may have had difficulty dealing with a particular exercise if he or she had never encountered a similar situation before. However, the level of performance in that particular exercise would not be so poor as to unduly affect the candidate's overall rating in that skill as reflected by assessor consensus. It would affect the variance and range of the scores obtained in the distribution of assessment center ratings.
obtained in the distribution of assessment center ratings.

In the job performance ratings the skills were derived from observations made over two to four years. This period is not unlike the process for determining the final consensus skill rating. The "highs" and "lows" of a candidate's skill demonstration in actual job performance tend to be averaged out over time. A supervisor takes all the demonstrations of skill levels in particular types of situations into account in arriving at a skill score. This would contribute to a restriction of range of these scores and appears to be a likely explanation of the apparent under-prediction and over-prediction phenomena found in a detailed examination of the predictor and criterion data.

**Skill Ratings**

There is a logic to the pattern of significant correlations obtained for the individual skill ratings. Six of the significant (or nearly significant) correlations were observed relative to skill performance in the two least ambiguous, strongest simulation exercises, the Inbasket and the Leaderless Group Discussion (LGD). There is less opportunity for variance of behavior in these two assessment exercises and in the job performance situations used as criterion measurements. In other words, there are a limited number of effective ways of completing an Inbasket in an assessment center and of completing administrative duties on the job. It may be easier for the assessor to measure the candidate's performance in the two relatively concrete exercises, the Inbasket and LGD, as compared to the more ambiguous Employee Counseling Exercise. Although the
Individual Problem Analysis taken as a whole was not significantly correlated with the job performance criterion of "situational analysis," two of the skills measured in this exercise were significantly correlated with their job performance counterparts. These two skills, communications and decisiveness, were very important skills in this exercise and were also two of the easier skills to measure.

**Exercise Ratings**

The two significant correlations obtained for the administrative and group categories indicated a moderate, but definite relationship between the assessment center ratings and the performance ratings in these two situations. Whatever was being measured in the Inbasket and Leaderless Group Discussion is related to what was measured by managers when rating candidates on their demonstrated skills in administrative and group interaction situations, respectively. The criterion measurements in these two cases were clearly defined and fairly limited to specific duties or situations.

In the case of the other two situations, situational analysis and individual interactions, the correlations were very low. As disturbing as these results may be, there are several possible explanations for the lack of correlation.

One possible reason could have been that the criteria for these two job situations may have been defined too broadly. In evaluating situational analysis, supervisors were asked to consider an incumbent's skills in dealing with a diverse group of individuals in a wide variety of situations. Do the criteria include too much to make
it relevant to the assessment center rating? A better question would be: is the assessment center trying to predict too much using an exercise with too little breadth? The assessment exercise was designed to measure a candidate's organizational, analytical and decision making skills. Perhaps the exercise needs to be redesigned or expanded to better measure the skills and behaviors important to a broader sampling of critical job functions. Further investigation is warranted to determine whether the problem lies with the assessment exercise or the performance rating criteria.

The same consideration may be applied to the relationship between the Employee Counseling Exercise and individual interaction on the job. The assessment center exercise attempted to tap the candidates' skills as demonstrated in a counseling/leadership situation with subordinates. The criterion measurement, however, consisted of ratings of skills demonstrated in interaction with people from different managerial levels, and involved wider ranges of topics and situations. The candidates may have been rated on the job according to how well they "got along with others" while the Employee Counseling Exercise focused more closely on their counseling and leadership effectiveness. Perhaps the assessment exercise was a good measurement of superior-subordinate relationships, but too limited an instrument to predict performance in other one-on-one situations.

Another explanation of the near-zero correlation between the Employee Counseling Exercise scores and individual interaction scores obtained on the job is that the raters of on-the-job performance could not effectively measure this criterion. It may have simply been very
difficult to observe how an incumbent counsels or interacts with other individuals.

A third explanation is that the assessment center exercise may not sample the actual content of the job. One manager in the Air Traffic Service offered the explanation that very little actual counseling occurs on the job even though chiefs will report that it does. Instead, chiefs are judged by management on how well they can "stand up to the union." It may be that the job task analysis and the content of the exercise should be reviewed and revised.

A final alternative, which the data support, is that the predictor measurement was invalid since responses in this exercise were "faked." Most of the assessment center candidates had previously attended the FAA's introduction to supervision course at the Management Training School. The portion of this course that teaches the supervisors how to counsel subordinates uses simulation exercises similar to the assessment center's Employee Counseling Exercise. Thus, many of the candidates have received very specific coaching in the appropriate techniques for handling a counseling situation. When the contamination that might have resulted from information being shared by candidates during breaks between assessment center exercises is also considered, it is evident that many of the candidates may have been "play-acting" a role for the assessment center. The false level of skill demonstrated in the simulation exercise might not have been sustained in their job performance. This would explain the lower means found for performance ratings as compared to means for the assessment center ratings in
this column as seen in Table 4. The correlations in this exercise are a reversal of the pattern of the relationship found between all the other exercises and job situations.

The most likely explanation for these correlations is that all the above factors were acting to a certain extent to influence the correlation. Further research is called for to determine whether the main problem lies with the predictor measure or the criterion measure.

**Concensus Skill Ratings**

The significant correlations found for six of the seven concensus skills are encouraging. It appears that there is a definite, albeit moderate, positive relationship between the assessment center as a predictor and the criterion of job performance as judged by supervisors. The notable exception to this is the total lack of correlation relative to the skill of judgment. Judgment is perhaps the most difficult of the seven skills to measure in either the predictor or the criterion environments. Whatever is being measured in the assessment center for the skill of judgment bears no relationship to what is being rated by performance measures on the job. This may be because judgment is the least quantifiable of the skills. The rating results may be too dependent upon who applies them. Further investigation is definitely warranted to determine exactly what is being measured and whether it is appropriate and job related.
Overall Ratings

As predicted, a moderate, positive correlation ($r = .33$) was reported for the relationship between assessment center overall scores and job performance overall scores. This moderate correlation most likely underrepresents the true correlation between assessment center ratings and performance ratings. Had a random distribution of candidates been used rather than the selected sample, the correlation would have probably been much higher. The multiple-hurdles approach used in the selection system to screen candidates reduced the sample size and thus attenuated the correlation coefficient. Statistical corrections for restriction of range would likely yield a higher correlation coefficient.

General Conclusions

It can be safely concluded that the FAA's Air Traffic assessment center for facility chiefs has indeed identified good managers. Out of 63 managers in the sample, only 14 received overall job performance scores that were lower than their overall assessment center scores. In about one-third of these cases the subjects had been at the facility for just over a year and were still establishing themselves in the job. Only one of the 63 subjects was rated as performing at a less than satisfactory level on the job.

The study does not and cannot draw any conclusions concerning the job performance of those persons who did not make the cut-off in the assessment center. These candidates were not considered further and according to the national guidelines for the air traffic selection system, they could not be selected for target positions. There is
no way of knowing if the assessment center process "missed" any good candidates since their performance can never be measured on the job.

This study, as in many earlier ones, was subject to many limiting factors such as restriction of range, small sample size and possible predictor or criterion invalidity. There is no way to control for the restriction of range in an operational environment since management is not willing to qualify or select those candidates that do not meet the selection criteria. Restriction of range also occurs when management "eliminates" many of the candidates during pre-screening. This was unfortunately an economic necessity for this particular program. The sample size was limited by the number of candidates presently in the job for which data were available. Although the use of supervisory appraisals of job performance as a criterion has been criticized in the literature, there is still a good, practical reason for employing them: management's opinion of the candidates produced by a selection system will make or break the program.

How does the job performance appraisal format and process in this study differ from those typically used to validate assessment center performance? In this study the raters considered an incumbent's skill performance in four specific job-related functional areas. The ratings were made by three to six people who supervised or had fairly close working knowledge of the incumbent's performance. One or more of the raters in each region had been trained as an assessor. This fact, coupled with the knowledge that the ratings were for research purposes only should have helped control some of the "halo" and
leniency effects that typically have impacted on performance appraisals.

The ratings for this study were being collected at the same time the regions were gathering information for the regular annual performance appraisals, so that the raters had objective job performance information for each candidate readily available. The ratings might still have been subject to some unsystematic and arbitrary observations made by the evaluators. However, though the raters initially made individual observations and ratings, the final rating was based on a pooled consensus drawn from all of the ratings.

What of the question of criterion contamination? Although the raters were informed that the job performance measure was being taken to assist in the validation of the assessment center, this author does not think that this contributed to the problem of criterion contamination. No rater had access to the assessment center data, and there was no discussion concerning a candidate's assessment center performance 2 to 4 years earlier. Although there may have been a tendency for management to remember which candidates in the region did not do well in the assessment center process, the results or scores of successful candidates who are eventually selected for positions are hardly remembered. Management cared only that the candidates had been "blessed" by the process and were qualified to be selected.

The main reason for validating selection procedures is to determine if measurements obtained via a predictor instrument agree with measurements gathered via a criterion instrument. The results of this study generally indicated that the air traffic assessment center
is indeed measuring the skills that make for successful facility chiefs.

Further investigation and research is called for to determine why the Employee Counseling Exercise and Individual Problem Analysis showed such low correlations with supervisory evaluations of incumbent job performance. The content of these exercises must be carefully compared with the content of the actual job. Further research is also warranted to determine why the skill of judgment as measured in the assessment center had no correlation to the job performance rating of that skill.

The Federal Aviation Administration is currently reviewing the economic feasibility of retaining the assessment center process as part of the selection system for Air Traffic GS-15 chiefs and deputy chiefs. This study will assist management in evaluating the utility of the assessment center. The moderate correlation of $r = .33$ which described the relationship between overall performance in the assessment center and overall job performance is actually much better than it appears due to the highly selected sample used in this study. The research results have clearly shown that the assessment center contributes valuable information to a total selection system for managers in the FAA Air Traffic Service.
Appendix A

The FAA Assessment Center

Within the Federal government the FAA has been one of the early pioneers of the assessment center method for selection and development. The agency unquestionably holds the record within government for the broadest application of the process for selection in a myriad of positions.

The agency first used the assessment center process in 1972 as part of a comprehensive selection system to identify GS-15 managers for an Executive Development Program. The positive results of that assessment center, plus strong support from high-level management convinced of the efficiency of the technique, led to a broadened application of the method throughout the agency.

In 1975 the FAA and a team of managerial consultants conducted a job task and skills analysis of GS-15 Air Traffic Control Center and Terminal Facility Chief and Deputy Chief positions with the idea of using the information to develop a comprehensive selection system for these positions. Interviews were conducted with 30 facility chiefs and deputies to determine what tasks they performed and with what frequency of occurrence. The chiefs and deputies were also questioned about the skills required to perform effectively in the position. The data gathered from the interviews and a rating questionnaire yielded a ranked list of skills, knowledge and abilities necessary for
effective performance in the positions. Information was collected that described in detail the most important and most frequent activities, duties and responsibilities performed by a facility chief or deputy chief. Using this information the FAA Air Traffic Service and the Office of Personnel and Training contracted with the U.S. Department of Agriculture and a private consulting firm to utilize simulation exercises to appraise candidates' potential to perform in the positions of GS-15 Air Traffic Facility Chiefs and Deputy Chiefs. The results of the program were viewed positively by management and a foundation was laid for its further use. The Office of Personnel and Training and the Air Traffic Service continued to work together to develop a comprehensive, multiple criteria, national selection system for GS-15 Air Route Control Center and Tower facility chiefs and deputy chiefs.

Although it would have been ideal to measure each candidate's performance on multiple instruments or criteria and to integrate the results for a "total look" at a candidate, this was not possible. Due to the large number of candidates nominating themselves for consideration, and too tight time and budget constraints, it was necessary to develop a "multiple hurdles" approach to the selection system. A four step selection process was developed which included a review of the candidate's previous employment experience and performance, a special supervisory appraisal of demonstrated performance in several critical skill areas, an assessment center and a panel interview with key management officials. Each important skill, knowledge and ability as determined by the job task skills
Step One: The Local Evaluation

A national announcement opened a 21-day bidding period for the selection system. Each candidate submitted an application form (Civil Service Commission Form SF-171) describing his or her work history and experience in air traffic and government to the personnel office in the candidate's region. Each of the 14 regional personnel offices evaluated the applications according to standardized criteria in a published rating guide. The candidates were awarded points based on their previous job experience, training, performance evaluations, education and awards. The weighting factors and the cutoff score had been determined by the joint decision of personnel and air traffic management. Candidates were expected to have performed in a minimum number of different positions in the formal air traffic career progression system. It was not necessary to have held every position on the career ladder, but candidates with broader backgrounds and a high level of geographic and functional mobility gained a greater number of points. Such candidates would ultimately rank higher on the final register when all criteria points were combined. Candidates had to score a minimum of 100 points to qualify for further consideration.

Step Two: The Division Level Review

The second step of the selection process was a written supervisory appraisal of the candidate's demonstrated performance in the knowledge and skill areas of adaptability, initiative,
dependability, motivation and leadership, and technical knowledge of air traffic and the National Airspace System. The performance appraisal was usually made by the candidate's first level supervisor and then forwarded to a panel of second level and third level supervisors (branch and division chiefs) for approval. No point value was associated with the appraisal. The result was simply referral or non-referral to the third step of the selection process. In actuality, very few candidates were eliminated from further consideration at this point. Regional management was rarely held accountable for its referral decisions and generally preferred to let the third step of the selection process, the assessment center, "do the dirty work" of screening out the candidates.

Judging from their poor performance in the subsequent assessment center, 15 to 20 percent of the candidates should have not been referred. The reasons for referral to the assessment center were usually weak. The regions did not have the intestinal fortitude to refuse candidates, or the regional management thought that the candidates were "borderline" and gave them the benefit of the doubt. The regions also used the assessment center to gain developmental information on candidates that were deemed to have potential for future consideration. Using the assessment center for developmental purposes was very costly to the agency. The high number of candidates failing the 1976 and 1977 assessment centers forced the regions to tighten their recommendation and referral process.

Step Three: The Assessment Center

The third step of the selection process, the FAA assessment
center, is the focus of this study. The first agency assessment center for Air Traffic GS-15 Facility Chiefs and Deputy Chiefs was conducted in the Fall of 1976 in three different geographic locations. The assessment center and all of the instruments used were developed by Assessment Designs, Inc. of Orlando, Florida. The firm also conducted the 3-day training session for agency assessors. Four exercises, a Leaderless Group Discussion, and Inbasket, an Individual Problem Analysis and an Employee Counseling Exercise were used to assess 60 candidates.

In spite of its cumbersome beginnings, management was pleased with the initial results of the assessment center. The enormity of effort that goes into conducting an assessment center, i.e., the time and expense and human resources, certainly contributed to the attention the process received. Management actively sought ways to reduce these cost factors without changing the integrity of the assessor product.

Most of these changes were implemented the following year, 1977. The assessment center was conducted once each in 1977, 1978, and 1979. After each of the assessment centers, management conducted reviews of the process with the expressed goal of increasing both the content and face validity of the assessment center. The agency was quick to recognize the importance of standardizing the assessment center process to minimize the differential treatment candidates might otherwise receive during the program and to insure that results from year-to-year would be comparable. Considerable effort was expended in
standardizing and equalizing the process for all the candidates involved. The following describes some of the changes that were made to increase standardization.

Program design and administration. The FAA's Executive Personnel Staff has administered the program from its inception. This author, the personnel psychologist, has been responsible for all aspects of program format and modification. She has been responsible for all assessor training and assessment center administration from 1977 to the present.

Location of the assessment center. The first and foremost change in the program was the centralization of the location of the assessment center to the FAA's Management Training School in Lawton, Oklahoma. The physical layout of this facility is ideal for conducting an assessment center. A large classroom is available, as are multiple smaller rooms for assessor training and later use for the Leaderless Group Discussions. The beds can be removed from the tenth floor of the dormitory so that the small but comfortable dormitory rooms can be used for candidate workrooms and assessor interview and exercise rooms.

Assessors and assessor training. Twenty to twenty-five agency assessors were used in the application of the assessment center each year. Each region contributed managers to the process, emphasizing the national aspect of the program. Three of the four assessors on each team were GS-15 air traffic managers who occupied target positions or higher (facility or regional branch chief positions). One of every four team members was a branch-chief-level representative
from a regional personnel or civil rights office. The teams were balanced as best as possible with regards to the different variables of regional representation, minority representation and assessor experience. A ratio of three new assessors to one previously trained assessor on each team fostered continuity of the program and contributed greatly to the quality of training. Great effort was made to insure that no candidate was assessed by anyone that knew the candidate in any way or had otherwise formed a previous opinion as to the candidate's abilities or reputation. This was very difficult in certain cases. Some of the candidates had traveled widely during their careers, or else their reputations had traveled for them. Every effort was made to assign the "famous" candidates to teams that knew them the least. Steps were taken to ensure that these candidates would not have to interact in one-on-one situations with assessors who knew them.

The assessors arrived on-site one week prior to the candidates. The assessors were housed in a hotel to ensure that they would have minimum opportunity to interact with the candidates staying in the dorm. Assessor training was conducted by the agency personnel psychologist. The formal training lasted for four ten-hour days. The assessors "took" the exercises themselves, practiced on each other, and then observed and interacted with six "sample candidates" performing the exercises. The sample candidates were volunteers from the regions who had chosen not to bid that particular year but wanted exposure to the assessment process and an opportunity to gain developmental feedback. The sample candidates were highly motivated
and their performance was very typical of actual candidates.

Each assessors "training team" had a chance to observe and write-up the four different exercise reports on four of the sample candidates. The entire assessor group observed each of the sample candidates in the employee counseling situations. This allowed the assessors a chance to see six different approaches to the problem as well as six different assessors as role players. Reports were written during the day and in the evening. Each assessor's report was read and critiqued by the entire assessor group. The group as a whole developed and modified the standards for performance provided with the exercises by the consultants who developed them. Informal and formal measurements of scoring reliability were taken for the entire group. Each training team member rated his or her colleagues in toughness and objectivity. The staff psychologist balanced each final assessor team for leniency and toughness as well as for the previously mentioned factors.

Candidates: Each candidate received a letter from the Executive Personnel Staff which explained the assessment center process and described the types of exercises they would be taking and the skills that the assessors would evaluate. When the candidates arrived they were given a letter of welcome and instructions to report to a briefing the following morning. The oral briefing once again reviewed the types of exercises and skills observed. The candidates were strongly encouraged to ask questions. The candidates were provided individual schedules. Although the exercises were presented to each candidate in the same order, candidate breaks occurred at different
times and for different lengths of time. However, no break lasted longer than two hours. Although tight time frames did not allow the use of a "warm up" exercise, all candidates started off with the same exercise, the Leaderless Group Discussion. This, to some extent, equalized the candidates' exposure and sensitization to the process as noted by Cohen (1978).

Candidates were responsible for picking up their individual exercise material and for taking it to a workroom. At the end of the allotted work time they collected their material and took it with them to their interview rooms. The importance of maintaining the time schedule was impressed upon them. Very few candidates lost track of time or failed to show up at the correct time. The administrative staff was very conscientious in starting and stopping the candidates on time to insure no unfair advantages would result.

Instructions. Most of the instructions given to the candidates were in written form to prevent ambiguity and ensure fairness. The written instructions were clear and concise and therefore required little interpretation by the candidate. Candidates were encouraged to ask questions for clarification purposes. At the same time, the assessment center administrator was very careful as to how the questions were answered. The candidates were, in addition to all other things, being evaluated on their perceptual and analytical ability—no small part of which is the ability to read, understand and follow instructions.

Pre-knowledge of exercises. It was almost impossible to control for a candidate's previous knowledge of the content of the assessment
center exercises. The assessment center administrator had to make some assumptions regarding previous knowledge. Several regions may have attempted to prepare their candidates by conducting extensive discussions of the assessment process or by evaluating the candidates in mock exercises to give them performance feedback. No solid evidence of this could be found. However, the practice was rumored to have occurred—but always in "another region." The only way to view an uncontrollable situation is to view it philosophically. One region's idea of "cheating" was in fact another region's idea of "developing and training."

Some problems may also have resulted because so many candidates required assessment. It was necessary one year to conduct three cycles, which took a total of ten days. No matter how much it is discouraged, it is a recognized fact among assessment center administrators that candidates discuss their assessment center experience with other candidates at the center and at their respective work sites upon their return. To the extent that precise or factual subject matter or parameters of the assessment center are revealed differentially among candidates, contaminating effects may occur (Cohen, 1978). Cohen recommends that a possible solution to this problem may be the use of similar, but different, exercises for each cycle or application. However, only once, during a four cycle assessment center, did management decide to change portions of the Inbasket and the Employee Counseling Exercise.
Another area of concern to management was the fact that there was no limit to the number of times a candidate could apply for the program. Consequently, approximately 25% of each year's candidates were repeaters. A very small proportion repeated the process more than twice. The population of repeaters could easily be divided into two groups: (a) those who did not make the cut-off in their first exposure to the assessment center but, after specific feedback and development, passed the second time, and (b) those who consistently continued to score below the cut-off point.

Role playing: Role playing was essential to performance in two of the exercises contained in the assessment center, the Employee Counseling Exercise and the Individual Problem Analysis. The four person assessor team was split into two role-playing partnerships. Both problems were designed to allow one or both of the assessors in the partnership to play the roles required in each exercise. It was felt that the assessors, after their training, would be well-qualified to provide standardized stimuli and elicit specific behaviors from the candidates. In any interview, the quality of the assessor/interviewer's performance greatly determined the candidate's effectiveness in demonstrating his or her skills and abilities. All assessors received the same training and amount of practice in these role situations. Written outlines and guides which included specific questions were provided each assessor. In addition, each received coaching from the psychologist as well as constructive and critical feedback from his or her fellow assessors. The role players always worked with the same partners in each application of the assessment
process. This model contributed to the consistency of the role playing and the reliability of the results.

Skills measured. The four assessment center simulation exercises measured the following seven skills:

1. **Perception and Analytical Ability.** The ability to identify, incorporate and comprehend the critical elements of a situation, to interpret implications of alternative courses of action, and to evaluate factors essential to a problem's solution. The ability to seek out pertinent data and put it together to solve a problem either with others or alone.

2. **Organizing and Planning.** The ability to establish courses of action for one's self and/or others in order to accomplish specific goals; to make effective use of personnel and other resources; to establish objectives and priorities.

3. **Leadership.** The ability to direct, motivate, develop, and coordinate the activities of others; to gain the respect and confidence of others; to delegate authority and direct the efforts of others toward a task solution.

4. **Decisiveness.** The ability to make decisions, render judgments, take action, make commitments, and support decisions when challenged.

5. ** Judgment.** The ability to use sound judgment to reach logical conclusions and make decisions concerning the use of resources, guidelines and stated policies; the ability to determine courses of actions, and define solutions to problems based on the evidence at hand.
6. **Interpersonal.** The ability to respond to and be sensitive to the needs, feelings, and capabilities of others; the ability to deal effectively with others, regardless of their status or position in both favorable and unfavorable situations; the ability to accept interpersonal differences.

7. **Oral and written communication.** The ability to effectively and clearly present and express information orally and in writing.

**Numerical ratings.** The skills measured in each exercise were scored on a seven-point scale according to standards set by the material and in training. The scale was defined as follows:

- **7** - Outstanding
- **6** - Well above satisfactory (a very strong performance)
- **5** - More than satisfactory (better than average performance)
- **4** - Satisfactory
- **3** - Less than satisfactory (somewhat weak performance)
- **2** - Well below satisfactory (a weak performance)
- **1** - Very weak performance

**Assessment center exercise design.** Each year a new set of exercises that were parallel, but not identical, to those of previous years was developed by Assessment Designs, Inc. The exercises were taken and reviewed by previously trained assessors months before actual training began. It was not unusual for the assessors to disagree with the consultant's "recommended solutions or actions" in the material. Once consensus on a more effective or more accurate approach was reached by the assessors during training, changes were
made in the assessor guide. Most often the problems were minor and due to the difference between government and private sector environments. This is mentioned only to emphasize the importance of checking and testing all simulation material, even if it has been customed designed for content validity and applicability to a particular position. Although the job task and skills analysis was used as the basis for developing the exercises, and although a great effort was made to ensure that the exercise environment was very realistic and sampled the content of the target position, occasionally small inconsistencies had to be corrected. Small things can frequently make a large difference in a candidate's perception of the entire process and possibly affect his or her motivation to perform as he or she would in "real life." It is very difficult for a candidate to be placed in an unfamiliar environment, to interact with new people and be required to perform effectively under simulated conditions within tight timeframes, knowing all the while that his or her career is on the line. Thus, the staff and the consultant sought out and utilized candidates and assessor feedback from previous assessment centers to increase the content validity or "real lifeness" of each succeeding assessment center. The goal was to accurately sample the domain of situations confronting a facility chief or deputy without duplicating the job activities identically.

**Skill definitions by exercise.** All candidates and assessors were provided with the same summary definition of each of the skills to be observed in the assessment center. It is important to note that each skill was demonstrated and measured in a different way in each
exercise. Thus, the skills were defined by each exercise's requirements and the behaviors were elicited in the particular context of each exercise. This emphasizes the importance of accurately defining and sampling the particular content of the skills as they are demonstrated in the target positions. Even though a particular skill may be measured in all four exercises, it may be emphasized and/or require tailoring in different ways from one exercise to another. The skills received different weightings across different exercises. The consultant provided extensive examples of what behavior to look for under each skill category in each exercise. This list was expanded upon by assessors during training.

**Assessment center exercise description.** The following provides a typical description of the four exercises used in the assessment center. They were a Leaderless Group Discussion, an Inbasket, an Individual Problem Analysis and an Employee Counseling Exercise.

In the **Leaderless Group Discussion** the candidates formed a committee charged with rank-ordering possible future sites of a regional office for a new government agency. Each candidate was assigned a particular geographic location which he or she was to advocate. The candidates were provided a common packet of information and specific data concerning their own individual sites. The exclusive information was used by each candidate in his or her presentation to the entire group. The candidate was expected to perceive and analyze the positive and negative aspects of both the common and the exclusive individual site information. The candidate had two tasks: (a) to advocate his or her assigned viewpoint, and
(b) to aid and facilitate the group in making the best decision for regional ranking. This simulation exercise tapped a candidate's skills in leading and facilitating a peer-group discussion. The candidate's efforts made to organize the material of the individual presentation as well as those efforts made to organize the group discussion and decision process were evaluated by the assessors. The skills emphasized in this exercise were leadership, communication, and interpersonal skills.

In the Inbasket Exercise the candidate assumed the role of a newly appointed director of a federal transportation agency district office. Time pressure was created by establishing a scenario in which the candidate had to catch a plane in a short time for a three-day trip. The candidate was given 2 1/2 hours to deal with a collection of forty memos, to become aware of certain situations existing in the new office, to organize the material, to plan future actions and to take action or direct others to take action on any pressing matters. This administrative exercise emphasized the skills of perception and analytical ability, organization and planning, decisiveness and judgment.

In the Individual Problem Analysis the candidate was requested by an urban affairs council to review a body of information concerning the feasibility of installing a mass transportation system. The candidate was expected to organize and analyze the data, choose between two alternative systems and present a recommendation both in writing and in an oral report to two assessors playing the roles of mayor and city council member. This was essentially an analysis and
staffing exercise which required the candidate to deal with a database and present or sell a decision to supervisors. Organization and planning skills were emphasized in this exercise.

The **Employee Counseling Exercise** placed the candidate in the role of a new office chief whose deputy would soon retire. The deputy had not dealt very effectively in the past with two disagreeing supervisors. The candidate was to meet with them to discuss their differences and lack of communication. This exercise put the candidate into a supervisory situation requiring him or her to probe for information, facilitate dialogue and lead and counsel two subordinates. The skills most heavily emphasized were leadership, communication, and interpersonal skills.

**Observing, writing, rating, and consensus.** The assessor training provided a great deal of practice in making observations of behavior, evaluating and discussing the behavior observed, and classifying, rating and scoring the candidate's performance against the standardized criteria. During the actual assessment center each assessor, working alone or with a partner, was assigned to observe each of six candidates as they performed in the different exercises. Each assessor was responsible for writing up six exercise reports. No assessor wrote two exercise reports for the same candidate. Each candidate participated in four exercises and was observed by four different assessors. During the consensus discussion on the third day each assessor would read his or her particular report for the candidate under discussion. Each of the four assessors recorded what exercise was being read. All four assessors then agreed on a
concensus score for each skill in each exercise. Each skill was examined across all four exercises to determine the overall concensus score for that skill. The concensus skill scores added to produce the assessment center total score. The personnel psychologist served as an arbitrator of differences and a breaker of tie votes. She also attempted to control for "halo effect" by cautioning the assessors not to confuse a candidate's performance in one exercise with that in another. An example of the assessment center performance rating matrix is shown in Figure 1.

Scores and cut-offs. The concensus scores for each skill could range from 1 to 7, from "very weak" to "outstanding." Scores tended to cluster in the 3 to 6 range, with the mean at 4, or "satisfactory." The assessment center total score could range from 7 to 49. Most candidates scored between 25 and 32 with the mean usually being 30, or just above "satisfactory."

In 1977 a strict cut-off of 28 points was established as governing who would be referred to the National Review Board Interview which was the fourth step of the selection process. The cut-off point was based on the premise that a satisfactory concensus scores of 4 in each of the seven skills sum to an overall total of 28. Thus a candidate could be satisfactory (4) in five of the skills and less than satisfactory (3) in a sixth skill, but still score 28 by having a more-than-satisfactory (5) rating in the seventh skill.

In 1979 the strict cut-off rule was modified. A task force of previously trained assessors recommended that a more flexible "window" system be used to refer candidates. Candidates scoring 32 points or
### Figure 1

**ASSESSMENT CENTER SCORE MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>INBASKET</th>
<th>INDIVIDUAL PROBLEM</th>
<th>EMPLOYEE COUNSELING</th>
<th>LEADERLESS GROUP DISC.</th>
<th>CONSENSUS SKILL RATING</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PERCEPTIVE &amp; ANALYTICAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ORGANIZING &amp; PLANNING</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LEADERSHIP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DECISIVENESS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>JUDGMENT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>INTERPERSONAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>COMMUNICATION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EXERCISE TOTALS:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>OVERALL TOTAL</strong></td>
</tr>
</tbody>
</table>

**CANDIDATE:** ____________________________  **REGION:** ____________________________
above were automatically referred to the fourth step in the selection process. The assessment center report on each candidate scoring from 25 to 31 was reviewed by that candidate's respective regional division chief before he or she could be referred to the National Review Board Interview. All candidates scoring 28 to 31 were referred, and one-half of the candidates scoring in the "less than satisfactory overall" range of 25 to 27 were also referred.

**Feedback.** In both the 1977 and 1978 assessment centers the candidates returned to their respective regions as soon as the last exercise was concluded. They received notification within five days as to whether they had passed or not passed. The handwritten narrative assessment center reports were completed on site, and typed at the Washington, D.C. headquarters of the FAA. The reports were mailed to the candidates through their respective air traffic division chiefs. The regions arranged for each candidate to receive extensive feedback from an assessor from that region who had participated in the assessment center. Usually six weeks elapsed before the candidate received specific information concerning his or her performance. The length of time between performance in the assessment center and receipt of feedback often resulted in a candidate's "amnesia" relative to his or her performance in the center. Many candidates who had failed could not recall taking certain actions in specific exercises or making statements that were recorded by the assessor. Resentment and hard feelings toward the process, as well as a lack of confidence in the process, began to develop. Clearly, a new procedure for giving feedback was necessary.
A task force comprised of regional air traffic managers recommended to management officials that each candidate be retained on site one extra day to receive feedback directly from a member of the team that had assessed him or her. This approach to the feedback session was positively received by management. The candidates spent one-half day developing solutions and recommendations for the Air Traffic Service concerning actual problems that were occurring in the field facilities. This time allowed the assessors to conduct the exercise and consensus discussions for each candidate's feedback. The candidates' reaction to learning immediately whether they had passed the process and to learning how they were perceived by the assessors was also positive. After the candidates departed, assessors had the remainder of the afternoon to complete the candidates final reports. The reports were collected, brought to FAA headquarters, typed and mailed to the respective regions. Each candidate's immediate supervisor arranged several discussion or counseling sessions to review the candidate's performance and to arrange for developmental training assignments. This method of feedback was much more acceptable to the candidates and virtually eliminated the problem of "amnesia" mentioned earlier.

Step Four: the National Review Board Interview

Candidates passing the assessment center were referred to the fourth step of the selection process, the National Review Board Interview conducted at FAA headquarters. The interview panel was comprised of headquarters and regional division chiefs, and representatives from the Offices of Civil Rights and Personnel and
Training. The interview was chaired by the Air Traffic Service Director. Each candidate spent anywhere from 1 to 1 1/2 hours in a basically unstructured interview setting answering questions and discussing a wide range of topics. Among the subjects covered were the technical aspects of air traffic control, approaches to managing an air traffic facility, labor-management relations, the National Airspace System and other national air traffic programs.

Candidates who were successful in the National Review Board Interview were placed on a national selection register for a 3-year period. Each region used the list to select candidates to fill its chief and deputy chief vacancies.
Appendix B

The Job Performance Appraisal

Instruction to the supervisors. The chief of the Executive Personnel Staff and the personnel psychologist sent a letter to each regional air traffic division chief. The letter briefly outlined the history and design of the national selection system for air traffic chiefs. The concept of a criterion oriented validation study was explained. Enclosed with the letter was the special rating form and instructions. The raters were assured that the performance data would be confidential and used for research purposes only. Each letter was followed by a phone call from the personnel psychologist to answer any questions of the division chiefs or staff members who would be primarily responsible for coordinating the rating process.

Performance rating process. The regional air traffic staffs convened a panel of branch chiefs and specialists to discuss each candidate's performance. Many of the people doing the performance ratings had been trained as assessors or at least had participated in the assessment center process as a candidate. A portion of the instructions the raters received were as follows:

In the assessment center, each candidate was given a score for each skill in each exercise. All the scores for one particular skill contributed to the "concensus score" for that skill. The concensus skill scores were added to produce the "total score" for the assessment center. The same format will be
used here to rate the incumbent's skills that were rated in "on the job" performance. The same skills that were rated in the assessment center are to be rated in the performance evaluation. However, instead of observing the candidate's performance in simulation exercises, you are to rate the incumbent's skills as you perceive they are evidenced in certain situations on the job.

The assessment center exercises were designed to sample the content of real-life situations a facility chief or deputy chief encounters on the job. The four assessment center simulation exercises required the candidate to demonstrate each skill in different ways. The performance evaluation form is divided into four different important functional areas of a facility chief or deputy chief position: the administrative area, the information processing and staff/situational analysis area, and the areas governing individual or one-on-one interactions and interactions within groups. Please try to evaluate each candidate as carefully as possible for each skill in each situation. Please also provide a summary consensus score for each skill. The consensus scores should be added to indicate a candidate's total score.

The raters were given a list of the skills to be rated and their definitions. The skills were the same as those observed in the assessment center (see Appendix A). The raters were also provided a copy of the seven-point rating guide scale (see Appendix A) as well as instructions on discussing skills and reaching consensus. In addition, the raters were provided the following description of the four
functional/situational job performance areas for which candidates were to be rated:

**Administrative.** Consider the incumbent's skills as demonstrated in the following situations: everyday paperwork processing data and information collection, correspondence, and communications as each relate to facility concerns, traffic count, personnel and budget matters, union issues, coordination of agency programs and relationships with users (pilots, airlines, airports, etc.) and the public. The assessment center exercise that attempted to measure administrative abilities was the Inbasket exercise.

**Situational analysis.** Consider the incumbent's skills in dealing with any situation and the people involved, and to make a decision and follow through with a recommendation or action. "On the job" examples may include situations involving facility technical problems, or facility or regional personnel problems, as well as situations involving input from people outside the facility such as airport personnel, media persons, consumer groups, union personnel, or political representatives. The assessment center exercise that attempted to measure a candidate's ability in this area was the Individual Problem Analysis.

**Individual interaction.** Consider the incumbent's skills and abilities when dealing with one-on-one relationships with supervisors, subordinates and peers. Consider the incumbent's relationships with individuals in the facility concerning technical matters, policy communication, personal interaction and employee counseling. The assessment center exercise that
attempted to measure individual interactions was the Employee Counseling Exercise.

Group interaction. Consider the incumbent's skills as demonstrated in the following situations: facility chief's conferences, regional and facility staff meetings, local coordination meetings, meetings with airport management or military groups, and special projects and task forces. The assessment center exercise that attempted to measure skills in group situations was the Leaderless Group Discussion.

An example of the performance matrix used for the "on-the-job" evaluations is shown in Figure 2. From three to six raters individually scored each candidate. They then shared their impressions and ratings with other raters and the entire group reached a consensus as to the scores to be placed on the final performance matrix. The Air Traffic Division Chief reviewed and approved all the ratings.
### Figure 2

#### CANDIDATE PERFORMANCE RATING MATRIX

<table>
<thead>
<tr>
<th></th>
<th>Administrative</th>
<th>Situational Analysis</th>
<th>Individual Interaction</th>
<th>Group Interaction</th>
<th>Consensus Skill Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perceptive &amp; Analytical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Organizing &amp; Planning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Leadership</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Decisiveness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Judgment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interpersonal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Communication</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exercise Totals:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Candidate:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Region:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C
Table 2
The Means, Standard Deviations, and Standard Errors of the Means for the Relationship Between the Inbasket and Administrative Functions

<table>
<thead>
<tr>
<th>Skills</th>
<th>Assessment Center</th>
<th>Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>5.15</td>
</tr>
<tr>
<td>Perceptive/Analytical</td>
<td>4.18</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>1.38</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>4.60</td>
<td>5.06</td>
</tr>
<tr>
<td></td>
<td>1.23</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>.16</td>
<td>.12</td>
</tr>
<tr>
<td>Leadership</td>
<td>3.95</td>
<td>5.24</td>
</tr>
<tr>
<td></td>
<td>1.23</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>.16</td>
<td>.11</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>4.16</td>
<td>5.32</td>
</tr>
<tr>
<td></td>
<td>1.31</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>.17</td>
<td>.11</td>
</tr>
<tr>
<td>Judgment</td>
<td>4.01</td>
<td>5.06</td>
</tr>
<tr>
<td></td>
<td>1.19</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>.15</td>
<td>.12</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>4.64</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>1.09</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>.14</td>
<td>.11</td>
</tr>
<tr>
<td>Communication</td>
<td>4.91</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.11</td>
<td>.13</td>
</tr>
<tr>
<td>Overall Exercise Correlation</td>
<td>30.8</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>6.6</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>.83</td>
<td>.67</td>
</tr>
</tbody>
</table>
Table 3

The Means, Standard Deviations, and Standard Errors of the Means for the Relationship Between the Individual Problem Analysis and Situational Analysis

<table>
<thead>
<tr>
<th>Skills</th>
<th>Assessment Center</th>
<th>Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>4.29</td>
</tr>
<tr>
<td>Perceptive/Analytical</td>
<td>Standard Dev.</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.17</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>Mean</td>
<td>4.57</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.15</td>
</tr>
<tr>
<td>Leadership</td>
<td>Mean</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>NA</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>Mean</td>
<td>4.58</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.15</td>
</tr>
<tr>
<td>Judgment</td>
<td>Mean</td>
<td>4.21</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.16</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>Mean</td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.11</td>
</tr>
<tr>
<td>Communication</td>
<td>Mean</td>
<td>4.81</td>
</tr>
<tr>
<td></td>
<td>Standard Dev.</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.12</td>
</tr>
<tr>
<td>Overall Exercise</td>
<td>Mean</td>
<td>31.63</td>
</tr>
<tr>
<td>Correlation</td>
<td>Standard Dev.</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>S.E.M.</td>
<td>.73</td>
</tr>
</tbody>
</table>
Table 4
The Means, Standard Deviations, and Standard Errors of the Means for the Relationship Between the Employee Counseling Exercise and Individual Interaction

<table>
<thead>
<tr>
<th>Skills</th>
<th>Assessment Center</th>
<th>Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Dev.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.E.M.</td>
</tr>
<tr>
<td>Perceptive/Analytical</td>
<td>4.82</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>.13</td>
<td>.12</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>5.00</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>.14</td>
<td>.11</td>
</tr>
<tr>
<td>Leadership</td>
<td>5.05</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>.13</td>
<td>.13</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>5.22</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>Judgment</td>
<td>4.69</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>.14</td>
<td>.14</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>4.95</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>.14</td>
<td>.14</td>
</tr>
<tr>
<td>Communication</td>
<td>5.13</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>Overall Exercise</td>
<td>34.9</td>
<td>5.5</td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td>.69</td>
</tr>
</tbody>
</table>
Table 5

The Means, Standard Deviations, and Standard Errors of the Means for the Relationship Between the Leaderless Group Discussion and Group Interaction

<table>
<thead>
<tr>
<th>Skills</th>
<th>Assessment Center</th>
<th>Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptive/Analytical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.91</td>
<td>5.03</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>1.02</td>
<td>.82</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.13</td>
<td>.10</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.65</td>
<td>4.93</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>.93</td>
<td>.76</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.12</td>
<td>.09</td>
</tr>
<tr>
<td>Leadership</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.25</td>
<td>4.80</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>1.07</td>
<td>.97</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.13</td>
<td>.12</td>
</tr>
<tr>
<td>Decisiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.59</td>
<td>5.03</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>.96</td>
<td>.82</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.12</td>
<td>.10</td>
</tr>
<tr>
<td>Judgment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.41</td>
<td>4.80</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>1.00</td>
<td>.88</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.13</td>
<td>.11</td>
</tr>
<tr>
<td>Interpersonal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.57</td>
<td>4.86</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>.78</td>
<td>.85</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.10</td>
<td>.11</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.83</td>
<td>4.83</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>.90</td>
<td>.90</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>Overall Exercise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>32.2</td>
<td>34.4</td>
</tr>
<tr>
<td>Standard Dev.</td>
<td>5.1</td>
<td>4.9</td>
</tr>
<tr>
<td>S.E.M.</td>
<td>.65</td>
<td>.61</td>
</tr>
</tbody>
</table>
Table 6

The Means, Standard Deviations, and Standard Errors of the Means for Relationship Between the Conensus Skill Ratings

<table>
<thead>
<tr>
<th>Skills</th>
<th>Assessment Center</th>
<th>Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptive/Analytical</td>
<td>Mean 4.42</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .88</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .11</td>
<td>.09</td>
</tr>
<tr>
<td>Organizing/Planning</td>
<td>Mean 4.68</td>
<td>4.93</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .80</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .10</td>
<td>.09</td>
</tr>
<tr>
<td>Leadership</td>
<td>Mean 4.40</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .68</td>
<td>.84</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .09</td>
<td>.11</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>Mean 4.57</td>
<td>5.15</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .65</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .08</td>
<td>.10</td>
</tr>
<tr>
<td>Judgment</td>
<td>Mean 4.24</td>
<td>4.85</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .71</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .09</td>
<td>.11</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>Mean 4.66</td>
<td>4.82</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .74</td>
<td>.80</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .09</td>
<td>.10</td>
</tr>
<tr>
<td>Communication</td>
<td>Mean 4.87</td>
<td>4.91</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. .75</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .09</td>
<td>.10</td>
</tr>
<tr>
<td>Overall Correlation</td>
<td>Mean 31.78</td>
<td>34.7</td>
</tr>
<tr>
<td></td>
<td>Standard Dev. 3.8</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>S.E.M. .48</td>
<td>.59</td>
</tr>
</tbody>
</table>
References


