THE GENERAL MANAGEMENT IN-BASKET (GMIB)

Technical Report

This report is an update of earlier papers on the General Management In-Basket, including:

* The item-by-item scored General Management In-Basket, International Personnel Management Association Assessment Council, Philadelphia, PA, 1987.

* The General Management In-Basket, 19th International Congress on the Assessment Center Method, Toronto, Canada, 1991.

Richard C. Joines

September 2011

MANAGEMENT & PERSONNEL SYSTEMS, INC.
www.mps-corp.com

The General Management In-Basket (GMIB)

The General Management In-Basket is the only in-basket test that has been reviewed for Buros' **Mental Measurements Yearbook**. The GMIB was included in the 12th edition, published in 1995.

In keeping with standard practice, two experts in the field independently reviewed the GMIB. Excerpts of their reviews are given below.

Buros Mental Measurements Yearbook
12th Edition, 1995

Reviewer #1

In summary, the General Management In-Basket is a sound test of management skills...The results will be informative to the candidates, and their acceptance of the test should be favorable..."

Reviewer #2

"In comparison to assessment centers, the GMIB provides a similar level of validity at significantly lower cost... In short, the GMIB should be seriously examined by any organization interested in improving the identification and selection of its management talent..."

MSAT FOR STUDENTS

In addition to the full GMIB test, a short form of the GMIB has been in usage for many years. In 2007, a new version of the GMIB Short Form was developed for students. It is known as the MSAT (Managerial Skills Assessment Test). The MSAT was implemented by one of the top ten business schools in the country and is now used as part of their educational curriculum for all business school students. As of 2011, over 5000 students have been tested. The MSAT is used to diagnose the student's current ability to lead and manage people. Each student is given a Personal Managerial Development Report. In addition, courses have been redesigned to provide coverage of the managerial concepts embedded in the MSAT.

THE GENERAL MANAGEMENT IN-BASKET

Overview

The GMIB was developed by the author in 1984. The first paper on the GMIB was presented at the 1987 International Personnel Management Association Assessment Council. This paper gave an analysis of problems associated with the scoring of the traditional in-basket, along with how these problems were solved by utilizing the GMIB item-by-item scoring methodology. Additional papers describing reliability and validity studies were presented at the 1989 and 1991 meetings of the International Congress on the Assessment Center Method.

This report is essentially a combination of the earlier papers, along with the results of additional studies that have been conducted. The first major validity study on the GMIB was conducted in 1987. The research method utilized in that study, along with the results, is described in detail in this report. Subsequent studies of the GMIB utilized essentially the same methodology, with the exception of a study conducted by the U.S. Department of Army. The results of these additional studies are briefly summarized in this report.

Substantial information on the performance of racial and sex groups is also available at this time and is described in this report. Finally, information pertaining to the relationship of the GMIB to other assessment instruments is summarized.

At the time of its development, the GMIB represented a significant departure from traditional in-basket technology. The approaches used in developing and scoring the GMIB were not based on any previously described methods or work (published or unpublished). The GMIB approach was original, designed specifically to overcome the more serious problems associated with in-basket testing: namely, problems related to time and difficulty of scoring, and scoring reliability.

One additional problem associated with the traditional approach, especially as used in the public sector, was the practice of emphasizing face validity in the in-basket scenario and items. The rationale for this approach pertained to concerns for content validity. Prior to the GMIB, in-baskets used in the public sector were typically developed for a specific position, with face validity maximized; and consequently, such in-baskets were only appropriate for the one position for which they were developed.

The GMIB departs from the approach of emphasizing face validity in the in-basket scenario and items. The GMIB places candidates in a neutral scenario and achieves content validity that generalizes across organizations and different management assignments. This was accomplished by constructing in-basket items that would be relevant across virtually all management jobs. A twofold strategy was used: (1) the GMIB was designed to be a "theory-based" in-basket, with many items oriented towards testing the application of management theory to practice; and (2) common management situations that

generalize across management positions were identified and items constructed to sample these situations.

While departing from traditional in-basket testing in significant ways, the GMIB nevertheless retains the key features of traditional in-basket testing that have generally been regarded as unique, valuable features of the in-basket testing format. Principal among these features is the requirement for candidates to formulate their own courses of action in response to the in-basket items, and write memos and letters as appropriate. Thus, the GMIB has not overcome traditional in-basket problems by merely substituting a closed-end (or multiple-choice) response format. Rather, the GMIB has overcome traditional in-basket problems by implementing a number of important changes that, when integrated, achieve the desired goals of greatly reduced scoring time, high reliability of scoring, and content validity that generalizes across different organizations and management jobs.

The utility of the GMIB approach has been demonstrated in a number of studies to be described in this report. These studies demonstrate that: (a) the inter-rater reliability of scoring the GMIB is typically on the order of .90 or higher; (b) the GMIB possesses significant, substantial criterion-related validity in predicting managerial success across different occupational groups; and (c) the GMIB has less adverse impact than reported in the literature for ability and achievement tests.

The following are key innovations introduced by the GMIB:

- Theory-based model used for development of leadership, employee motivation, morale and empowerment items, including theories pertaining to participative leadership, motivation-hygiene (intrinsic vs extrinsic job satisfaction factors), and situational leadership

- Items developed to be relevant across organizations, different management jobs, different assignments, and level of supervision

- All power items; no simplistic "throw-away" items; traditional in-baskets frequently included items such as magazine articles that should be read when time allowed and/or routed to others and/or other simplistic items that did not involve any substantive knowledge or skill area that warranted independent measurement

- Separate answer forms, thus no need to review test pages for notes; one answer form designed to capture candidate's actions and understanding of management issues presented by the item, another answer form for use in writing memos and letters

- Elimination of in-basket features and scoring methods that are readily susceptible to training or coaching; "problem analysis" is <u>not</u> scored by observing whether candidates "tie" items together; "planning & organizing" is <u>not</u> scored by observing whether candidates prioritize or handle the more important items first, or whether candidates spend time dealing with "throw-away" items

- Items scored independently on 0-4 or 0-5 basis; critical items scored on five point scale with one point allotted to candidates who fully attempt the item, regardless of item response (introduced as a means of indirectly measuring "planning & organizing" skill)

- Total score arrived at by simply summing item scores, i.e., no need to try to relate information to dimensions to arrive at in-basket total score

- Only one rater required due to extremely high inter-rater reliability resulting from new scoring approach (second rater does not add significantly to scoring reliability due to high single rater reliability)

- Factor analysis results used to derive "dimension" scores; the four factors measured by the GMIB are (1) Leadership Style and Practices; (2) Handling Priorities and Sensitive Situations; (3) Managing Conflict; and (4) Organizational Practices/Management Control

- Automated, yet individually tailored candidate feedback reports with suggested developmental needs and learning objectives for each of the four factors measured; additionally, reports include feedback on speed of processing administrative work based on number of items completed (a feature not included in traditional in-baskets)

In addition to the standard public and private sector executive versions of the GMIB, forms are also available for all Public Safety supervisory level positions (Police Sergeant through Chief; and Fire Captain through Chief). Special forms of the executive version are available for engineers and school administrators. The time allowed candidates to complete the public and private sector executive forms of the GMIB is two hours and forty-five minutes. The law enforcement and fire versions allow candidates three hours.

GMIB candidate scores can be reported against a variety of norms. Scores may be reported relative to the entire data base, or alternately, scores may be reported relative to candidates at the same level of supervision as those tested (e.g., candidates for a first-level supervisory job may be compared to other candidates who tested for a first-level supervisory job). Additionally, scores may be reported against norms maintained for certain specialized groups, including police, fire and engineers. Over 20,000 candidates in the United States and Canada have taken one version or another of the GMIB.

The GMIB can be taken in paper/pencil format, or preferably, online. When the online version is used, the test appears in one window, and the response form in another. In either version, the requirements are the same: candidates must play the role of a manager, analyze the issues that come to their attention in their "in-basket," decide what actions to take, and write memos or letters or send email as they deem appropriate in handling their in-basket.

Problems Associated with the Traditional In-Basket

Traditional Scoring Methodology

In-baskets have traditionally been scored on assessment dimensions, including skills such as problem analysis, judgment and decision making, interpersonal sensitivity, planning and organizing, management control and written communications. A team of two or three assessors typically review and evaluate a particular in-basket. This evaluation process is often aided by having the candidates complete a form to briefly explain their actions. A form is sometimes included that requires candidates to indicate the order in which they completed the items; or alternately, a form is sometimes included that requires candidates to indicate their views on the relative importance, or priority, of the items.

In addition, many assessment centers are also designed to include an in-basket interview to further explore each candidate's underlying reasons for handling the items as indicated by the completed work. The traditional approach to scoring in-baskets is consistent with the standard assessment center process of classifying observed behaviors into assessment dimensions, then assigning a score (usually 1 - 5) on each dimension considered relevant to the assessment exercise.

Within the general schema of scoring in-baskets on "dimensions," a number of different approaches to scoring in-baskets may be found in the literature, the goal of which appears to be to simplify the scoring process and/or make it more reliable. The various approaches to scoring in-baskets have been categorized as either "content" or "stylistic" schemes (Schippmann et. al., 1990). Content schemes are highly objective and consist of a counting of characteristics of the assessee's in-basket, such as the number of decisions made, memos written, etc. Stylistic schemes are based on assessor ratings of the assessee's degree of possession of an individual difference variable. This approach requires assessors to evaluate the quality of the assessee's work, such as the quality of decisions made as opposed to simply counting the number of decisions made.

Regardless of the approach taken, the traditional approach involves arriving at a numerical score on each dimension measured by the in-basket. In arriving at this dimension score, some researchers may combine the content and stylistic schemes. In addition, an overall in-basket score is sometimes reported and this score may be derived in a number of ways, ranging from some combination of dimension scores to a review of all available information by a panel of assessors leading to a subjectively assigned overall score.

Another way of viewing the variety of approaches used to score in-basket dimensions is on a continuum ranging from holistic to objective. Holistic scoring would fall into the "content" category. In its purest form, assessors would simply review all work completed by the candidate, identify those actions and decisions (i.e., behaviors) that are considered to pertain to a specific assessment dimension, then assign a rating that best describes the assessee's degree of possession of the assessment dimension. This approach relies upon the assessor's training and judgment to identify all available information relevant to each dimension, then make a qualitative evaluation about the assessee on each dimension. The holistic approach is consistent with the standard assessment center method for scoring assessment exercises, whether they are in written or oral format.

It is not contrary to the holistic approach to provide guidance to the assessors on the items contained in the in-basket, as long as the assessors are still given the responsibility of combining their findings across items to assign scores on the dimensions. A common way to develop such guidance is to convene a group of Subject Matter Experts (SME's), usually supervisors of the target position, to review the in-basket items. During the review process, the SME's generate a list of possible actions that might be taken in effectively handling each item. The assessors utilize this guidance in forming their own judgments of the candidate's overall skill on a particular dimension by reviewing the candidate's work across all items; using the preestablished guidance as an aid only.

Objective scoring, in its most extreme form, consists of the previously described, simplistic "counting" rules (i.e., number of decisions made), with dimension ratings resulting from the various objective counts that are made of the various elements of the candidate's overall performance in the in-basket. Such an approach effectively removes the assessor's judgment in favor of the decision rules which are adopted. These purely objective approaches appear completely contrary to the standard assessment center methodology of relying upon assessors to evaluate the quality of a candidate's behavior and the degree to which the candidate possesses the various assessment dimensions (i.e., holistic approach).

Some approaches have attempted to combine the concern for quality with some objective approach to scoring in-basket dimensions. These approaches typically reduce the role of the assessor's judgment in assigning final dimension scores while permitting assessor judgment to determine the quality of the candidate's behavior with regard to performance on the in-basket items and/or on the dimensions with regard to particular in-basket items. The overall score on dimensions is derived through combining performance on the items. These approaches tend to increase reliability of scoring but may also restrict assessor judgment to the point that the validity of the in-basket is likely to suffer.

For example, Brannick, et. al. (1989) used SME's to develop a list of possible responses for each in-basket item and judged each response as positive, neutral or negative on each of five in-basket dimensions (organizing & planning, perceptiveness, delegation, leadership, and decision-making). Assessors used the guidance to assign candidates scores of +1, 0, or -1 on <u>each</u> dimension for <u>each</u> item. Such an approach appears designed to evaluate the quality of the candidate's responses while insuring simplicity and consistency in the assignment of dimension scores, thereby leading to high reliability of overall dimension scores. However, this approach is overly simplistic, resulting in reliability, but questionable validity.

In explanation of the scoring approach, the researchers stated:

> "...if an assessee failed to take action on a 'red hot' item, then the behavior was scored negatively for organizing and planning, leadership, perceptiveness, and decision making (p.960)".

While high total score reliability was obtained (r=.95), artificially inflated correlations between dimensions were caused by scoring the same behavior (i.e., item) on multiple dimensions. Since dimensions are designed to tap different elements of behavior, the scoring scheme in this study seriously erred by requiring a rating of quality on <u>each</u> dimension for <u>each</u> in-basket item.

By removing the necessity for assessors to struggle with the issues normally confronted by assessors (i.e., categorize the behavior into only the dimension or dimensions it represents; evaluate the behavior(s) pertinent to each dimension, etc.), the scoring procedure sacrificed those assessment principles which are the key elements of obtaining valid dimension ratings.

Traditional in-basket holistic scoring would first require that the behavior be classified into the assessment dimension that it represented prior to evaluating the "quality" of the behavior. Failure to deal with a "red hot" item would typically be considered a function of poor planning & organizing (especially if the candidate completed the in-basket items in the same order in which they were provided in the test, and failed to peruse the in-basket to find any "red hot" items). To conclude that failure to deal with a "red hot item" is a poor indication on most or all of the dimensions being assessed confounds the assessment dimensions and leads to spuriously high correlations among the various dimensions being rated (i.e., if someone misses a "red hot item," they would likely be rated low on all the dimensions, thus, the correlation among the dimensions would automatically be high).

Not surprisingly, the approach used by the researchers resulted in high reliability of scoring. For the reasons noted above, however, the validity of the dimension ratings was sacrificed. Interestingly, the researchers note the lack of divergent validity among dimensions -- but due to the scoring scheme adopted, this could hardly be otherwise.

The Brannick et. al. (1989) study is interesting in that it reflects the confusion that has surrounded in-basket scoring and attempts to improve inter-rater reliability in scoring in-baskets. The desire to attain high reliability has resulted in approaches that clearly lead to high inter-rater reliability, but only by adopting simplistic rules that lead to low construct validity of the dimensions.

The variety of approaches used to score in-baskets suggests a problem in itself. Practitioners and researchers are well aware of the difficulties in arriving at dimension scores when scoring in-baskets. It appears that these difficulties account for the wide variety of more objective schemes that have been attempted and reported in the literature. Objective approaches which attain high reliability but which either reduce the role of the assessor to simplistically counting the number or kind of actions taken, or which rigidly impose a decision rule such as generalizing a single positive behavior to represent positive standing on all dimensions being rated, appear to confuse the need for reliability with the need for validity.

Objective counting schemes, or combined quality/counting schemes, do not appear to be in favor with other kinds of simulations. For example, leaderless group discussions (LGDs) focus on the effectiveness of a candidate's oral communication skill. The present review did not reveal any LGD scoring schemes in which oral communications was measured by counting the number of words spoken by a candidate during an LGD or any similar simplistic scoring method.

Similarly, candidates in the LGD are typically evaluated on their leadership skill by the effectiveness of their participation and influence in guiding or directing the group to the attainment of its goals, not the mere amount of participation or number of attempts to lead or guide. The fact that a number of in-basket studies have used "counting" approaches, or combinations of quality/counting methods, suggests

significant dissatisfaction with standard assessment procedures when it comes to scoring in-baskets.

Difficulty of Scoring In-Baskets/Alternatives

The difficulty inherent in scoring traditional in-baskets is a function of many variables, including the length of the in-basket and its complexity. Assessors are confronted with a wealth of information and must attempt to keep track of everything the candidate did so that they may correctly relate this information to the dimensions being assessed. Assessors often experience serious cognitive overload and rating reliability suffers.

Due to scoring difficulties, many practitioners have been hesitant to use an in-basket when examining a large number of candidates. Recognizing these problems, Kraus (1986) reported on the development of a multiple choice in-basket, describing the issue as follows:

> "Large candidate populations usually preclude a test developer's use of examination modes such as orals, essays and assessment centers. This becomes acute when testing for middle-to-upper management positions, since those examination methodologies which are usually considered the least efficient are, in fact, often the most preferred."

Although the cost effectiveness and utility of a multiple choice in-basket is very appealing, a very important ingredient is missing in this approach. Lopez (1966) stressed the value of in-baskets in measuring recall rather than recognition. It would seem that the feature which most distinguishes in-baskets from multiple choice tests is the requirement to formulate responses to situations rather than selecting the best choice from among a given set of alternatives. Where the candidate's choices are restricted, there is no method by which to determine just how disastrous a candidate's approach to solving a problem might be -- or, for that matter, just how creative an approach the candidate might adopt.

Additionally, it may be argued that a multiple choice in-basket is not really an in-basket. Fidelity with the job situation is low since supervisors on the job actually formulate their responses to situations, write memos and letters, etc. They do not respond to a set of predetermined multiple choice alternatives in selecting a course of action. Therefore, so-called multiple choice in-baskets would more properly be considered multiple choice tests based on situational stimuli. The multiple choice approach does not solve the problems associated with scoring the traditional, open-ended response format utilized by in-baskets; it merely replaces the troublesome open-ended response format with the multiple choice format.

Replacing the traditional in-basket with a multiple choice test format is akin to developing a multiple choice format for a projective test. Would we still consider a test such as the Rorschach a projective test if we replaced the open-ended response format with a multiple choice format? Unlikely. It is because of the open-ended response format that we consider the Rorschach a projective test. Changing the response format represents a contradiction in test terminology; and the same holds true for the traditional in-basket. In short, we should not be talking about multiple choice projective tests, and we should not be talking about multiple choice in-basket tests.

Reliability and Validity

The research on the reliability and validity of in-baskets is mixed. Positive results were reported by Bray and Grant (1969). They found that the in-basket contributed significant validity over and above that obtained for paper and pencil tests.

Hinrichs and Haanpera (1976) evaluated the internal consistency reliability of various situational exercises and found reliabilities in the range of .22 to .41. These researchers concluded that the internal consistency reliability of situational exercises was lacking, and in particular, noted that the in-basket and the job environment report were the two exercises most in need of overhaul or replacement.

Kesselman, Lopez and Lopez (1982), in commenting on prior studies, expressed their view as follows:

> "A second reason, and the one which we hypothesize is largely responsible for the ambivalent results, concerns the methodology involved in ascertaining a participant's in-basket performance. Research related to interpersonal perception, including performance evaluation, the validity of the employment interview, and interpersonal attraction have consistently pointed out the distortion that inevitably occurs when one human being observes, recalls and evaluates the actions of another."

Based on these concerns, Kesselman et. al., conducted extensive research to develop an objective scoring procedure for an in-basket designed for first-level supervisors in a utility firm. This procedure is similar to the multiple choice format adopted by Kraus (1986), except that candidates first complete the in-basket, then complete a self-report questionnaire to indicate which of the listed actions they took when handling the item. The self-report questionnaire is then objectively scored.

The researchers reported a split-half, odd-even reliability coefficient of .83 and conclude that their work demonstrates that in-baskets can indeed be scored with satisfactory reliability using an objective scoring key. This was accomplished, however, by scoring self-report data and not the actual narrative in-basket responses of candidates.

In a review of the reliability and validity of in-baskets (Schippmann, et. al., 1990), the authors concluded that in-baskets can be reliably scored but that obtained reliability coefficients are "modest at best". With regard to inter-rater reliability, the authors included 13 studies in their review, with 10 studies reporting the obtained range of coefficients across rated dimensions and three studies reporting a single reliability coefficient for in-basket performance.

In an effort to obtain some rough quantification of the inter-rater reliability studies reviewed by Schippmann, et. al., the mean of the coefficients forming the low end of the range was calculated for the 10 studies reporting a range of reliabilities. The resultant mean reliability was .60. For the three studies reporting a single reliability coefficient, the mean reliability was found to be .77. Thus, the authors characterization of obtained reliability coefficients as "modest" appears warranted.

With regard to in-basket validity, Schippmann et. al. conclude that validity "is at best marginal." The studies included a wide-range of criterion measures, from indirect measures such as grades, scores on

standardized exams and salary progress to more direct measures such as supervisory ratings of on-the-job performance. For the more direct measures of performance, 13 studies report a range of validity coefficients for the in-basket dimensions.

The means of the validity coefficients forming the low and high ends of the reported ranges were calculated for the 13 studies in this category, resulting in a mean range of validity coefficients of -.25 to .36. The mean negative validity forming the low end of this range was not due to any single large negative validity coefficient, but rather, a host of negative values across the studies reviewed.

Based on this global analysis, the characterization of in-basket validity as being "marginal" also seems to be warranted. In short, the available research on the in-basket suggests that it is plagued by problems related to reliability and that these problems are not offset by high validity.

The traditional, holistic approach to scoring in-baskets is difficult and complex which no doubt is a major reason for the disappointing reliability results. On the other hand, simplistic "counting" approaches appear to sacrifice a key element of the assessment center method -- that of evaluating the quality of the candidate's behavior. While high reliability in overall score may be attained using such approaches, their is nothing in the literature to suggest higher validity or other favorable psychometric properties, such as divergent and convergent validity (i.e., construct validity).

There are two major problems preventing the reliable scoring of in-baskets using the traditional, holistic approach: (1) the pure cognitive demands placed upon assessors; and (2) the lack of standardization of the behavioral information base.

Gaugler and Thornton (1989) demonstrated the difficulties assessors experience in processing all of the information at their disposal. These researchers found that assessors have difficulties in properly classifying observed behaviors into their appropriate assessment dimensions, and that as the number of dimensions being evaluated increases, classification errors increase. Specifically, they found significant differences in classification accuracy when the number of dimensions being assessed increased from three to six; with even more error when the number of dimensions was increased from six to nine.

In-baskets have traditionally been designed to measure 5 - 7 dimensions of performance based on candidate responses to 20 - 30 in-basket items. The information processing demands inherent in such a task are complex and make the job of scoring an in-basket difficult and time consuming, but even with substantial time taken to score an in-basket, the task is so demanding that considerable error in the accuracy of behavior classification is likely to be present.

In addition to the pure information processing demands and the difficulties experienced by assessors in reliably classifying and making holistic evaluations of behavior on each assessment dimension, there is a lack of standardization of the behavioral information base. This issue has not been noted in the literature and no research has been conducted to investigate this issue.

However, if we examine the task of the assessor in scoring a traditional in-basket, we find that the

problem is not solely one of processing and recalling a large amount of information common to all assessees. The assessor's job is more complex than this because the same information on all candidates (i.e., standardized situation) is not available. Rather, the assessor must process quantitatively and qualitatively different information on all candidates with the goal of reliably rating all candidates on all dimensions.

In-baskets are typically designed to require candidates to effectively plan and organize their time. The time limits generally prevent a substantial number of the candidates from completing all items as a means of assessing whether candidates proceed in an organized manner, insuring that priority items are handled within the allotted time.

The candidates are free to choose the items they complete. Consequently, it is common for candidates to complete different items in the time allotted. In addition, the number of items completed may vary significantly. There are no accepted or standard rating principles to address all of these potential differences in arriving at ratings of candidate ability levels on the dimensions being assessed. Should a candidate who completes a small number of items but who uses good judgment on those items be rated higher on "judgment" than a candidate who completes all items but who demonstrates poor judgment in several instances? Could the second candidate have demonstrated judgment equal to the first candidate had he/she decided to only attempt a small number of items but to do them well?

The issue becomes even more complex since standardization does not even exist for the same item completed by two candidates. Since traditional in-baskets allow for managerial stylistic differences in scoring, there is typically no "one best method" for handling a particular item. The assumption is that several approaches may be equally effective. By scoring candidates on dimensions, it is assumed that patterns of strengths and weaknesses on the separate dimensions may be determined by evaluating the candidate's work across all items. The way in which a candidate handles a particular item determines the "dimension information" that may be derived from the handling of that item. This leads to a lack of standardization with regard to the dimension information produced by any given item.

For example, one candidate might choose to write a memo in response to an in-basket item. The content of the memo, along with the way in which the memo is written, may provide evaluative information on "judgment/decision making" and "interpersonal sensitivity," along with "skill in written communication." Another candidate, however, might choose to delegate action on the item, thereby providing evaluative information on "skill in delegation." Thus, the same item may provide information on completely different dimensions for the two candidates, leading to a lack of standardized information with regard to the dimensions included for measurement in the in-basket.

Given many such combinations across even a small number of in-baskets, raters find that they must rate dimensions such as problem analysis, judgment, management control, interpersonal sensitivity, etc., based on a different behavioral information base for each candidate. As the number of candidates increases, raters find it increasingly difficult to insure that they have rated all in-baskets using the same rating standards.

This lack of standardization with regard to the evaluative information available to assessors at the

"item" level is a function of the dimension scoring methodology. Dimension scoring, in turn, is an approach which assumes that greater validity is obtained in deriving information across items than by evaluating the appropriateness of action on each item and combining these results across items. It is this underlying assumption that requires examination.

The "standardization" problem would not exist if the candidate's appropriateness of action on each item were evaluated with no concern for determining the particular skills or abilities (i.e., dimensions) demonstrated by the candidate in handling the item. If this approach were used, a standardized way of evaluating candidate responses would exist and scoring should be more reliable. The reason this approach has probably not been used is because of devotion to the "dimension" scoring approach and the unique information on particular strengths and weaknesses that only dimension scoring is thought to offer.

Problems Resulting From Existing Assumptions about In-Baskets

The way in which in-baskets are currently used appear to embody two questionable underlying assumptions:

(1) The practice of tailoring in-basket exercises to specific positions reflects the assumption that the content validity of the process is based on building high face validity (i.e., the in-basket scenario must place the candidate in the same position as the target position and pose issues comparable to those encountered by incumbents of the target position).

(2) In-baskets must be scored on assessment dimensions in order to be valid in selection and worthwhile in terms of the developmental feedback information available to assessees.

Assumption #1 -- High Face Validity as a Condition of Content Validity

The widespread usage of in-baskets specially constructed to possess high face validity, especially in the public sector, reflects two prevailing beliefs: (1) that this is a necessary condition of content validity; and (2) that candidate acceptance will be low unless face validity is high.

The key issue in content validity, however, pertains to whether the types of supervisory or management situations encountered on the job are representatively sampled by the test. Face validity is not a condition of content validity, yet according to a relatively recent review of in-baskets (Schippmann, et. al., 1990, p. 851):

> "...in all the research reviewed, there was no evidence of any effort to develop a test plan or item budget for construction of in-basket tests following conventional procedures. For example, none of the studies attempted to obtain importance ratings of job-related tasks or knowledge, skills, and abilities (KSA's) and to use this information to guide in-basket construction in a way that would be consistent with content-oriented test development procedures as described by Schippmann, Hughes & Prien (1987)."

Current methods for constructing in-baskets are not clearly defined. To the extent that they are based on job analysis, most practitioners use an approach consisting of identifying tasks, grouping tasks into duty areas (or behavioral domains), identifying KSA's, and linking KSA's to tasks. If the practitioner attempts to utilize the job analysis information to develop an in-basket, the task information will be carefully reviewed. To the extent that meaningful duty areas have been identified, the practitioner will attempt to construct items that pertain to the duty areas -- provided it is realistic to incorporate such job tasks into the in-basket format.

Job analysis results depend upon the type of job analysis method chosen, and in particular, the way of organizing information about a job. If a position specific approach is used, and if task information is grouped into duty areas, the resultant job analysis information base may be thought of as "functional."

In this strategy, the task information is tailored to the specific position. "Manages or resolves conflicts among professional staff in order to maintain cooperative and productive working relationships," is typically refined (made narrower). In the case of an in-basket for a computer firm, for example, a position specific, functional approach might result in the task statement: "Manages or resolves conflicts among computer programmers with regard to new product development in order to maintain cooperative and productive working relationships."

In the functional approach, the task would likely be grouped into an identifiable area of job responsibility (i.e., job function), such as "New Product Development." The functional approach focuses on the specific goals of the work group or identifiable areas of responsibility. For a personnel manager, these might be areas such as Recruitment, Examining, Classification, Compensation, and Budgeting. Supervisory or individual tasks within each area would be listed.

This method of job analysis does <u>not</u> tend to group the tasks of a supervisor into <u>process</u> areas, yet the role of a supervisor or manager is largely to accomplish results through others, which is more of a process, or series of processes, than a series of specific job functions.

A "process-oriented" job analysis geared toward identifying the supervisory elements of a job would be more likely to group tasks into areas such as "managing conflict," "developing staff," "maintaining/improving employee motivation and morale," etc. Where defined in this manner, tasks become more general in nature. There is no need to tie a task to a specific organizational functional area; thus, for example, a task statement pertaining to managing conflict would cut across all potential functional areas of the job.

Further, the functional approach to job analysis is likely to overlook and fail to identify many critical process-oriented tasks. When SME's think in terms of a functional area (e.g., new product development), they are likely to identify the functionally-related tasks, not the process-oriented tasks that are independent of functional area. For this reason, functionally-oriented job analysis studies tend to result in examination plans which focus on functional areas of the job and which suggest the need for position specific tasks -- while overlooking the process-oriented nature of the position.

Where jobs are compared in terms of process, most supervisory positions will be found to contain

comparable "content" areas. As will soon be seen, the identification of common content "process" areas was instrumental in development of the GMIB.

Where content validity is "high" based upon the clear relevance of in-basket problems to major supervisory processes, candidate acceptance appears to be high, even where there is absolutely no face validity for a particular target position. Practitioners typically judge candidate acceptance by complaints or protests filed by candidates. As will be seen in the discussion of the GMIB, the problems are generic in nature, and designed to be applicable to any supervisory or managerial job.

The GMIB is not face valid for any particular job, yet clients routinely report positive feedback from candidates; purely testimonial evidence only, but nevertheless, an indicator of acceptance. More objective is the fact that out of the 20,000 or so candidates that have taken the GMIB for a large variety of job titles in both public and private organizations, only one complaint alleging a lack of job-relatedness has been filed (the matter was heard, and dismissed).

Assumption #2 - Dimension Scoring is Required

The second assumption that appears to have been embraced by assessment center practitioners is that in-baskets must be scored on specific assessment dimensions. In the research conducted by Kesselman et. al. (l982), however, the researchers state:

> "The high intercorrelations among the subscores supports previous in-basket research which suggests that the underlying ability measured by the in-basket exercise is a single generalized trait."

Other researchers have found little support for the construct validity of assessment center dimension ratings and have suggested that assessors tend to evaluate performance in exercises (known as the "exercise effect") rather than performance on dimensions (Sackett & Dreher, 1982; Sackett & Dreher, 1984).

Given problems associated with accurately classifying behaviors into dimensions, the time associated with scoring in-baskets, reliability and validity concerns, coupled with the likelihood that dimension ratings may not be serving their intended purpose of providing accuracy of information on the degree to which the candidate possesses the dimension (e.g., little evidence for construct validity of dimensions), the logical question would appear to be, "Why attempt to measure dimensions?"

Instead, why not measure the appropriateness of action on each in-basket item, then determine through factor analysis whether item performance provides reliable dimension (i.e., factor) information? At least this approach should lead to a standardized and more reliable scoring system. The validity of the in-basket as determined in empirical research would then be the appropriate way to determine the overall merits of the new approach.

As will be shown, the research evidence strongly supports this model.

Description of GMIB Item Types, Scoring and Reporting Systems and Research Results

In the private sector version, candidates assume the role of the Director of the XYZ Division of a hypothetical organization. In the government version, candidates play the role of the manager of the XYZ Division of the Department of Good City Government. The in-basket scenario is neutral and specifically designed to be unlike any known position or organization. In this sense, the GMIB has no face validity for any position. The GMIB consists of 15 items that deal with the following kinds of general supervisory/management issues or "process" areas:

* Employee motivation & morale
* Interpersonal conflicts
* Implementation of new procedures
* Dealing with personnel external to the organization
* Delegation; upward and downward
* Performance problems
* Staff development/growth
* Work organization/efficiency
* Group dynamics/team efforts

The process content areas were identified by reviewing the results of a large number of supervisory/managerial job analysis studies of different positions at different organizational levels, the goal of which was to serve as the basis for the development of assessment centers for selection. The above areas seemed to capture most of the kinds of generic process-oriented tasks that were common to all or most of the positions studied. Expert judgment alone was used in identifying these content areas.

The processes used by the manager create an atmosphere within the work group. The GMIB was designed to select managers who empower others using participative principles coupled with an understanding of employee motivation and morale, while still maintaining a commitment to excellence in work group outcomes. Examples of desirable, behavioral work group outcomes that may be associated with the processes used by effective supervisors and managers are as follows:

* Behavior teaches subordinates that they are expected to successfully perform and meet objectives

* Behavior teaches subordinates that they are adults and will be held accountable

* Behavior teaches subordinates that they are to make every effort to solve problems in an effective manner and to avoid attempts at casting the manager in the role of an autocrat who solves all problems

* Behavior teaches subordinates that their ideas and input are valued

* Behavior teaches subordinates that they are "team members" with a common goal of achieving high productivity and successful organizational outcomes

Candidate Response Format

For each item attempted, candidates respond on standardized forms, of which there are two. On the first, candidates are instructed to analyze the supervisory/managerial issues involved in the item, even if they believe the item requires no immediate action. Subsequently, and on the same form, candidates are instructed to describe any actions they would take in handling the item, either in the present or at a future date. Finally, a second form is supplied to candidates on which they are to write any memos or letters that they would write in handling the in-basket item. Each form is numbered to correspond to the item. Only the response forms are necessary to score the in-basket since all analyses and actions must be shown on these forms.

Scoring Method

The GMIB is scored on an item-by-item basis. Detailed rating guidance exists for each item. This guidance includes a narrative discussion explaining the goal of the item and an analysis of the management issues that are involved. In addition, rating scales anchored with descriptions of the responses to be rated at each level are used. Three of the 15 in-basket items are considered critical and these are scored on a scale ranging from 0 to 5. The remaining 12 items are scored on a scale ranging from 0 to 4. Candidates who "fully attempt" a critical item are awarded one point regardless of their analysis of the issues involved or the actions taken. This scoring method was used as a means of giving some credit to candidates who planned and organized their time sufficiently to ensure they dealt with the problems judged to be critical. It is worth noting at this point that research on the GMIB in predicting on-the-job ratings of "planning & organizing" skill has produced significant, substantial validity coefficients.

Based on factor analysis results, item scores are combined based on factor loadings to generate scores on the following four factors:

1. Leadership Style and Practices
2. Handling Priorities and Sensitive Situations
3. Managing Conflict
4. Organizational Practices/Management Control

Basis for Scoring Guidance

In an effort to devise an in-basket that would have validity across a wide range of management situations, a number of the items are geared toward the application of management theory to practice. In particular, concepts related to McGregor's Theory Y principles of participative management and Herzberg's Motivation-Hygiene Theory are utilized in the narrative scoring guidance. For items that have a theoretical orientation, the proper handling of the items is based more on acceptance of the theory and of its application to an applied situation than on the judgments of particular subject matter

experts or assessors.

Thus, the GMIB scoring guidance would not be valid in an organization that desired an autocratic style of leadership. In this sense, GMIB scores may be viewed as possessing "theory-based" validity. As an example of theory-based validity, suppose an organization was recently purchased and the new ownership found that all supervisors and managers were highly autocratic and only considered their autocratic subordinates to be successful performers. Any criterion-related validity study in this setting that utilized supervisory performance ratings as a criterion measure would find no validity or negative validity for an assessment procedure that led to high scores by highly skilled participative managers, even though we have 30+ years of research suggesting that participative managers are more likely to be successful over the long run.

As in all criterion-related studies, the results are severely limited by the available criterion measures. If supervisory ratings have no overlap with "true" success (i.e., ultimate criterion), then we would expect a test with true validity of 1.00 to yield an obtained validity of .00.

On the other hand, the new ownership could determine that it wanted supervisors who were successful participative, "team" leaders. To the extent that an assessment exercise correctly rank-ordered candidates on their overall skill as participative leaders, the test would be measuring what it was supposed to measure -- and would be valid. Thus, theory-based validity would exist to the extent that the test measured the ability of candidates to apply "participative" or "team" leadership theories to practical, applied situations.

Content validity is most directly present when a test simulates job performance. Content validity refers to a "method of measurement." In our example, content validity would be present to the extent that the test simulated realistic job situations. If these simulations are designed to measure application of a theory of leadership or employee motivation to applied situations, with scoring based on correct application of the theory, then the test may be thought of as possessing "theory-based, content validity." All that is necessary is that the situations be common to supervision and that the test results produce their intended result.

Not all items in the GMIB are theory-based. Some items in the in-basket are based on commonly accepted principles of organizational effectiveness and sound management practice. Items in these categories include issues such as dealing with a performance problem or responding to an important public official on a sensitive matter. However, even in these item types, the methods of handling the items that are viewed as superior are consistent with the underlying assumptions of McGregor's Theory Y and/or common organizational goals (e.g., need to be responsive, maintain positive image, etc.)

Reliability of Ratings

Nineteen inter-rater reliability studies have been conducted on the GMIB. A preliminary study, not listed below, was used to investigate and refine the scoring guidance ® = .81, 6 raters, 10 in-baskets). Since then, 19 studies have been conducted. The lowest obtained coefficient was .86. The simple average of the 42 obtained coefficients is .92. If the coefficients are weighted by the number of in-

baskets rated, the mean is .93. Table 1 summarizes the results of these studies.

Where inter-rater reliability is high, there are decreasing returns associated with adding additional raters. The Spearman-Brown formula may be used to estimate the reliability of the scoring process using two raters instead of one. Given a single rater reliability of .86, using two raters increases the reliability to .92. If the single rater reliability is .95, the reliability with two raters is .97. Thus, the improvement is negligible in both cases and does not warrant adding a second rater. Clearly, this has significant implications for savings in terms of time and costs.

Table 1

INTER-RATER RELIABILITY FOR GMIB TOTAL SCORE

	No. of Raters	# of In-Baskets	# of Coefficients	Mean r
Study #1	5	10	10	.86
Study #2	5	10	10	.94
Study #3	4	10	6	.93
Study #4	2	100	1	.95
Study #5	2	10	1	.94
Study #6	2	20	1	.86
Study #7	2	20	1	.89
Study #8	2	20	1	.93
Study #9	2	20	1	.95
Study #10	2	20	1	.95
Study #11	2	20	1	.94
Study #12	2	20	1	.92
Study #13	2	22	1	.92
Study #14	2	8	1	.95
Study #15	2	28	1	.91
Study #16	2	30	1	.94
Study #17	2	22	1	.96
Study #18	2	28	1	.93
Study #19	2	36	1	.97

Feedback Reports

Trained raters average approximately 15-30 minutes to thoroughly score the GMIB. This includes assigning scores on each item attempted by the candidate as well as selecting from among a bank of narrative statements descriptive of the candidate's performance on each item.

When the rater has completed the scoring form, the data is entered into a custom data base program. The GMIB data base program generates a bar chart profile on each candidate, showing the candidate's percentile standing on the test overall and on each of the four factors measured by the GMIB. The normative data used to generate bar chart profiles may vary, from using all candidates in the data base vs. subsets that may be created by specifying criteria (e.g., job type, organizational level of the candidates, type of organization, etc.).

The data base program outputs both score information and information on "why" the candidate received a certain score on a certain item (as determined by the trained GMIB rater). The "why" leads to the specific statements included in the candidate's evaluation/feedback report. The length of the report is a function of the number of items attempted by the candidate. Feedback is only given on the items attempted. Feedback consists of a skill description statement for each item attempted, organized according to the four factors measured by the GMIB. For each skill description statement, the candidate's associated developmental needs are also specified, along with suggested learning objectives. Based on the number of items attempted by the candidate, relative to the average number of items attempted by candidates in the data base, the system also generates a statement descriptive of the candidate's speed in processing administrative workload (i.e., below average, average, or above average).

Combining the item-by-item scoring approach with factor analysis made it possible to develop a candidate feedback reporting system that is tailored to each candidate's performance with regard to each test factor. Candidate feedback reports describe the candidate's skill level with regard to each GMIB item attempted within each test factor, along with associated developmental needs and specific learning objectives.

The reporting system was accomplished by developing an inventory of the ways in which different candidates handled each item. For each approach identified, a skill description statement was written along with the implied developmental needs and associated learning objectives. Approximately 1000 candidates were tested and "inventoried" before the reporting system reached a point of stability. There has been little need to change it since that time. The important point to note is that the GMIB feedback report is tailored to the candidate's actual performance on the test.

Original GMIB Criterion-Related Validation Study

Three hundred sixty-five employees of a public sector organization completed the GMIB as an initial hurdle in competition for selection into a management development program. The sample consisted of incumbents in levels 2, 3 and 4 of the organization's classification structure. Employees from approximately 120 separate job classifications were represented in the sample.

Level two applicants were non-supervisory higher-level professional personnel. Level three applicants were generally first level supervisors and level four applicants were either second or third level supervisors. There were 219 level two candidates, 102 level three candidates and 44 level four candidates.

Performance ratings were concurrently collected from immediate and next-higher-level supervisors. The number of completed performance rating forms was 278 for immediate supervisors and 243 for next-higher-level supervisors. Ratings on 194 subjects were available by both raters. This permitted an evaluation of the reliability of the criterion measures as well as the formation of several overall composite measures based on both sets of ratings.

Ratings were made on a nine point rating scale (1= low; 9 = high) on the following performance dimensions: (1) written communications; (2) leadership; (3) interpersonal relations; (4) planning and organizing; (5) analyzing problems and making sound decisions; and (6) oral communications.

For each performance dimension, two ratings were made: (1) the employee was rated in relation to employees at the "same" organizational level, and; (2) the employee was rated in relation to "all" employees at organizational levels 2, 3 and 4.

After rating employees on the performance dimensions, raters were asked to supply an overall rating of the employee in relation to employees at the same organizational level. The same nine point rating scale was used. This measure is hereinafter referred to as the "subjective" overall rating, one being made by the immediate supervisor and the other by the next-higher-level supervisor of the employee.

In addition to the subjective overall measure of performance, a series of mechanically derived overall measures of performance were formed. These measures were sums of the ratings made on the six dimensions by the immediate and next-higher-level supervisors, as follows:

1. Sum of immediate supervisor's ratings on the six performance dimension ratings using candidates at the "same" organizational level as the reference group.

2. Sum of next-higher-level supervisor's ratings for the "same" level reference group.

3. Sum of immediate supervisor's ratings on the six performance dimensions using "all" employees as the reference group (i.e., all employees at organization levels 2, 3 and 4).

4. Sum of next-higher-level supervisor's ratings on the six performance dimensions using "all" employees as the reference group.

Table 2 gives the means and standard deviations for the in-basket and various criterion measures, along with the sample size upon which each statistic is based.

Table 3 shows the obtained intercorrelations of the six performance dimensions based on ratings made using candidates at the "same" level as the reference group. The correlations below the diagonal are based on ratings by immediate supervisors (average n = 274). Correlations above the diagonal are based on ratings by next-higher-level supervisors (average n = 241).

Table 4 provides the intercorrelations based on "all" candidates as the reference group, with correlations below the diagonal based on ratings by immediate supervisors (average n = 276) and those above the diagonal based on next-higher-level supervisors (average n = 242).

Estimates of the reliability of the in-basket and the performance dimension ratings made by supervisors were obtained using Cronbach's (1951) coefficient alpha based on 99 cases for which complete data on all in-basket items and performance measures was available. These results are given in Table 5. Table 6 presents in-basket item correlations with total in-basket scores. These correlations (n = 365) ranged from .37 to .52.

Table 2

MEANS AND STANDARD DEVIATIONS OF IN-BASKET AND CRITERIA

	Mean	SD	N
Predictor			
In-Basket Total Score	18.14	8.44	365
Performance Measures			
Subjective Overall Rating by Immediate Supervisors	6.76	1.70	274
Subjective Overall Rating by Next-Higher-Level Supervisors	6.60	1.56	239
"Same" Level Ratings by Immediate Supervisors			
Written Communications	6.52	1.91	277
Leadership	6.39	1.83	274
Interpersonal Relations	6.60	1.88	274
Planning & Organizing	6.80	1.82	275
Analyzing Problems/Making Decisions	6.75	1.77	273
Oral Communications	6.68	1.81	276
Mean Dimension Rating	6.61	1.54	272
"All" Levels Ratings by Immediate Supervisors			
Written Communications	5.96	2.00	278
Leadership	5.86	1.93	276
Interpersonal Relations	6.28	1.92	276
Planning & Organizing	6.33	1.90	277
Analyzing Problems/Making Decisions	6.26	1.83	276
Oral Communications	6.27	1.79	278
Mean Dimension Rating	6.15	1.62	275
"Same" Level Ratings by Next-Higher-Level Supervisors			
Written Communications	6.34	1.78	240
Leadership	6.21	1.76	241
Interpersonal Relations	6.51	1.71	241
Planning & Organizing	6.60	1.67	242
Analyzing Problems/Making Decisions	6.52	1.72	242
Oral Communications	6.51	1.84	242
Mean Dimension Rating	6.44	1.49	238
"All" Levels Ratings By Next-Higher-Level Supervisors			
Written Communications	6.11	1.79	241
Leadership	5.88	1.84	242
Interpersonal Relations	6.37	1.71	243
Planning & Organizing	6.37	1.69	242
Analyzing Problems/Making Decisions	6.28	1.72	243
Oral Communications	6.34	1.72	243
Mean Dimension Rating	6.23	1.52	239

Table 3

INTERCORRELATIONS OF JOB PERFORMANCE RATINGS WITH "SAME LEVEL" EMPLOYEES AS REFERENCE GROUP: IMMEDIATE SUPERVISOR RESULTS BELOW DIAGONAL AND NEXT-HIGHER-LEVEL SUPERVISOR RESULTS ABOVE DIAGONAL

Dimensions	Written	Leadership	Interpersonal	P & O	Probs/Dec.	Oral
Written	- -	.72	.52	.73	.74	.75
Leadership	.62	- -	.69	.67	.75	.74
Interpersonal	.53	.68	- -	.53	.57	.62
P & O	.59	.68	.51	- -	.79	.70
Probs/Dec.	.64	.74	.61	.75	- -	.71
Oral	.72	.70	.71	.55	.67	- -

Table 4

INTERCORRELATIONS OF JOB PERFORMANCE RATINGS WITH "ALL LEVELS" AS REFERENCE GROUP: IMMEDIATE SUPERVISOR RESULTS BELOW DIAGONAL AND NEXT-HIGHER-LEVEL SUPERVISOR RESULTS ABOVE DIAGONAL

Dimensions	Written	Leadership	Interpersonal	P & O	Probs/Dec.	Oral
Written	- -	.70	.57	.78	.75	.76
Leadership	.68	- -	.69	.73	.74	.71
Interpersonal	.54	.68	- -	.58	.60	.65
P & O	.65	.77	.57	- -	.80	.71
Probs/Dec.	.69	.79	.65	.78	- -	.72
Oral	.70	.70	.69	.59	.69	- -

Table 5

COEFFICIENT ALPHA FOR PREDICTOR AND CRITERION MEASURES

Measures	Coefficient Alpha
Predictor	
In-Basket (15 Items)	.71
Criteria	
"Same" Level Ratings by Immediate Supervisors	.92
"All" Levels Ratings by Immediate Supervisors	.92
"Same" Level Ratings by Next-Higher-Level Supervisors	.91
"All" Levels Ratings by Next-Higher-Level Supervisors	.92

Table 6

IN-BASKET ITEM - TOTAL CORRELATIONS (N = 365)

Item Number	Pearson **r**
1	.51
2	.37
3	.47
4	.48
5	.42
6	.50
7	.46
8	.37
9	.39
10	.42
11	.41
12	.52
13	.43
14	.48
15	.46

Validity Coefficients: Table 7 gives the obtained validity coefficients for total in-basket score in

relation to the subjective "overall" ratings of performance and the mechanically derived "overall" measures. As will be noted, all validity coefficients are highly significant.

Table 8 gives the obtained validity coefficients for the six performance dimensions based on ratings by immediate and next-higher-level supervisors for the "same level" reference group. All validity coefficients are highly significant.

Table 9 gives the obtained validity coefficients for the six performance dimensions based on ratings by immediate and next-higher-level supervisors for the "all levels" reference group. Once again, all validity coefficients are highly significant.

Table 7

OBTAINED VALIDITY COEFFICIENTS: CORRELATION OF TOTAL IN-BASKET SCORE WITH OVERALL MEASURES OF PERFORMANCE BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS BASED ON "SAME" LEVEL AND "ALL" LEVELS REFERENCE GROUPS

Immediate Supervisors			Next-Higher-Level Supervisors		
Subj. Overall Rating	Same Level Mechanical Overall	All Levels Mechanical Overall	Subj. Overall Rating	Same Level Mechanical Overall	All Levels Mechanical Overall
r = .28* (n = 274)	r = .31* (n = 272)	r = .31* (n = 275)	r = .29* (n = 239)	r = .34* (n = 238)	r = .33* (n = 239)

* $p < .0001$

Table 8

OBTAINED VALIDITY COEFFICIENTS: CORRELATION OF TOTAL IN-BASKET SCORE WITH RATINGS MADE BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS ON SIX PERFORMANCE DIMENSIONS FOR "SAME LEVEL" REFERENCE GROUP

Performance Dimension	Immediate Supervisors	Next-Higher-Level Supervisors
Written Communications	.29* (n = 277)	.36* (n = 240)
Leadership	.26* (n = 274)	.30* (n = 241)
Interpersonal Relations	.19** (n = 274)	.18** (n = 241)
Planning & Organizing	.30* (n = 275)	.32* (n = 242)
Analyzing Problems/Decisions	.27* (n = 273)	.24* (n = 242)
Oral Communications	.27* (n = 276)	.31* (n = 242)

* $p < .0002$ ** $p < .006$

Table 9

OBTAINED VALIDITY COEFFICIENTS: CORRELATION OF TOTAL IN-BASKET SCORE WITH RATINGS MADE BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS ON SIX PERFORMANCE DIMENSIONS FOR "ALL LEVELS" REFERENCE GROUP

Performance Dimension	Immediate Supervisors	Next-Higher-Level Supervisors
Written Communications	.31* (n = 278)	.38* (n = 241)
Leadership	.26* (n = 276)	.23** (n = 242)
Interpersonal Relations	.21** (n = 276)	.19** (n = 243)
Planning & Organizing	.30* (n = 277)	.32* (n = 242)
Analyzing Problems/Decisions	.25* (n = 276)	.28* (n = 243)
Oral Communications	.29* (n = 278)	.31* (n = 243)

* $p < .0001$ ** $p < .005$

Reliability of Criterion Measures

In order to obtain an estimate of the reliability of the performance rating criterion measures, the ratings of immediate and next-higher-level supervisors were correlated, as follows:

(1) Subjective overall ratings of performance relative to those at the "same" level (yields reliability of subjective overall ratings)

(2) Sum of six dimension ratings for the "same" level reference group (yields reliability of the mechanically derived composite measure of overall performance for the "same" level ratings)

(3) Sum of six dimension ratings for the "all" levels reference group (yields reliability of the mechanically derived composite measure of overall performance for the "all" levels ratings)

(4) Ratings on each of the six performance dimensions for the "same" level reference group (yields reliability of ratings for each performance dimension on the "same" level ratings)

(5) Ratings on each of the six performance dimensions for the "all" levels reference group (yields reliability of ratings for each performance dimension on the "all" levels ratings)

With regard to the three overall measures of performance, reliabilities consistent with published research were found, ranging from .56 to .62. These results are given in Table 10. The results for the "same" level and "all" levels ratings on the six performance dimensions are given in Table 11.

Table 10

INTER-RATER RELIABILITY COEFFICIENTS FOR
THREE MEASURES OF OVERALL PERFORMANCE

Overall Measure of Performance	Inter-Rater Reliability
Subjective Overall Rating	.62 (n = 192)
Mechanically Derived “Same” Level	.61 (n = 192)
Mechanically Derived “All” Levels	.56 (n = 194)

Table 11

INTER-RATER RELIABILITY COEFFICIENTS FOR RATINGS BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS ON SIX PERFORMANCE DIMENSIONS

Performance Dimension	Inter-Rater Reliability	
	"Same" Level	"All" Levels
Written Communications	.51 (n = 194)	.46 (n = 196)
Leadership	.61 (n = 195)	.51 (n = 197)
Interpersonal Relations	.47 (n = 195)	.45 (n = 198)
Planning & Organizing	.52 (n = 196)	.50 (n = 197)
Analyzing Problems/Decisions	.50 (n = 196)	.50 (n = 198)
Oral Communications	.48 (n = 196)	.43 (n = 198)

The obtained validity coefficients were corrected for unreliability in the criterion in order to estimate true validity. Table 12 gives the corrected validity coefficients for the overall measures of performance. These validities are the best estimates available of the true validity of the GMIB in predicting overall job performance.

Table 12

BEST ESTIMATES OF TRUE VALIDITY: OBTAINED VALIDITIES CORRECTED FOR CRITERION UNRELIABILITY

Level of Supervisor Making Rating	Subjective Overall Rating	Mechanical Overall Rating "Same" Level	Mechanical Overall Rating "All" Levels
Immediate Supervisor	.35 (n = 274)	.40 (n = 272)	.41 (n = 275)
Next-Higher-Level Supervisor	.37 (n = 239)	.44 (n = 238)	.44 (n = 239)

Table 13 gives the corrected validity coefficients for the six performance dimension criteria (immediate and next-higher-level supervisors vs. "same" and "all" levels ratings). These estimates are the best estimates of the true validity of the GMIB in predicting performance within specific supervisory/managerial skill areas.

Table 13

BEST ESTIMATES OF TRUE VALIDITY: "SAME" AND "ALL" LEVELS
DIMENSION VALIDITIES CORRECTED FOR CRITERION UNRELIABILITY

Performance Dimension	"Same" Level		"All" Levels	
	Immediate	Next-Higher	Immediate	Next-Higher
Written Communications	.41	.50	.46	.56
Leadership	.33	.38	.36	.32
Interpersonal Relations	.28	.26	.31	.28
Planning & Organizing	.42	.44	.42	.45
Analyzing Problems/Decisions	.38	.34	.35	.40
Oral Communications	.39	.45	.44	.47

Additional Criterion Measures

The ratings made by immediate and next-higher-level supervisors were combined (sum or mean) to form the following measures of overall performance:

(1) Subjective Overall Combined Rating
(2) Mechanical Overall Combined Rating: Same Level
(3) Mechanical Overall Combined Rating: All Levels

In addition, the ratings of immediate and next-higher-level supervisors on each of the six performance dimensions relative to the "same" level reference group were summed.

A mean (or sum) of ratings by two raters will be a more reliable measure than either, provided there is reliability in the ratings. To estimate the reliability of the composite of two raters, the Spearman-Brown prophecy formula is typically used. Assuming interchangeability of raters, such an approach is comparable to doubling a test in length to increase its reliability.

Table 14 gives the obtained coefficients for the overall measures of performance, along with the reliability of each measure and the estimated true validity coefficient. Table 15 provides this information on the six performance dimensions.

Table 14

OBTAINED VALIDITY COEFFICIENTS, RELIABILITY AND ESTIMATED TRUE VALIDITY IN PREDICTING OVERALL MEASURES OF PERFORMANCE BASED ON COMBINED RATINGS OF IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Combined Rating	.28*	192	.76	.32
Mechanical Overall Combined Rating: Same Level	.34*	192	.76	.39
Mechanical Overall Combined Rating: All Levels	.31*	194	.71	.37

* $p < .0001$

Table 15

OBTAINED VALIDITY COEFFICIENTS, RELIABILITY AND ESTIMATED TRUE VALIDITY IN PREDICTING MEAN PERFORMANCE DIMENSION RATINGS FOR "SAME" LEVEL RATINGS BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS

Performance Dimension	Obtained Validity	Sample Size	Reliability	True Validity
Written Communications	.38*	194	.68	.46
Leadership	.29*	195	.76	.33
Interpersonal Relations	.18*	195	.64	.23
Planning & Organizing	.36*	196	.68	.44
Analyzing Problems/Decisions	.28*	196	.67	.34
Oral Communications	.28*	196	.65	.35

* $p < .0001$

Validity coefficients for the three organizational levels included in the study demonstrated similar patterns. Obtained and estimated true validity coefficients for the three overall measures based on the combined ratings of immediate and next-higher-level supervisors are shown in Table 16. Collectively, these results indicate that the validity of the GMIB is not restricted to any one organizational level (i.e., the GMIB is not "level" bound). Since the sample consisted of a wide variety of classifications (120 total), the GMIB also does not appear "position" bound; rather the results indicate that GMIB validity generalizes across positions and organizational levels.

Table 16

OBTAINED VALIDITY COEFFICIENTS, RELIABILITY AND ESTIMATED TRUE VALIDITY IN PREDICTING OVERALL MEASURES OF PERFORMANCE BASED ON COMBINED RATINGS OF IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS FOR EACH ORGANIZATIONAL LEVEL INCLUDED IN THE STUDY

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Level 2: Subjective Overall Combined Rating	.30*	185	.72	.35
Level 2: Mechanical Overall Combined Rating: Same Level	.27*	117	.71	.32
Level 2: Mechanical Overall Combined Rating: All Levels	.24*	118	.64	.30
Level 3: Subjective Overall Combined Rating	.33*	77	.86	.36
Level 3: Mechanical Overall Combined Rating: Same Level	.39*	60	.85	.42
Level 3: Mechanical Overall Combined Rating: All Levels	.34*	61	.79	.38
Level 4: Subjective Overall Combined Rating	.14	43	.76	.16
Level 4: Mechanical Overall Combined Rating: Same Level	.50**	19	.71	.59
Level 3: Mechanical Overall Combined Rating: All Levels	.44**	19	.62	.56

* $p < .005$ ** $p < .05$

Table 17 presents obtained validity coefficients for blacks, whites, males and females. The pattern of coefficients was not suggestive of differential validity and no formal technical analysis of regression line slopes or intercepts was indicated (there were no significant differences between black and white validity coefficients; and two of the three coefficients for females were higher than for males).

Table 17

OBTAINED VALIDITY COEFFICIENTS IN PREDICTING OVERALL MEASURES OF PERFORMANCE BASED ON COMBINED RATINGS OF IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS FOR RACIAL AND SEX GROUPS

Overall Criterion Measures	Obtained Validity Coefficients				
	Whites	Blacks	Hispanics	Males	Females
Subjective Overall Combined Rating	.21 (n = 208)	.18 (n = 38)	.39 (n = 52)	.25 (n = 135)	.37 (n = 186)
Mechanical Overall Combined Rating: Same Level	.26 (n = 128)	.31 (n = 27)	.42 (n = 33)	.30 (n = 79)	.37 (n = 117)
Mechanical Overall Combined Rating: All Levels	.24 (n = 130)	.20 (n = 27)	.46 (n = 33)	.35 (n = 79)	.28 (n = 119)

Factor Analysis

An exploratory factor analysis was performed on the scores of the 365 candidates in the sample. The intent was to determine whether independent and interpretable factors could be identified. Therefore, a principal components factor analysis was conducted using a varimax rotation (Kim, 1975). The Kaiser criterion of extracting only factors with an Eigenvalue greater than one was applied.

Four interpretable factors, accounting for 50% of the variance in total scores, were identified and named as shown below:

1. Leadership Style and Practices
2. Handling Priorities & Sensitive Situations
3. Managing Conflict
4. Organizational Practices/Management Control

Factor 1 (Eigenvalue = 3.12) clustered items dealing with an understanding of leadership and motivation principles, along with an understanding of how to vary the amount of direction given subordinates depending on the situation.

Factor 2 (Eigenvalue = 1.86) grouped together those items that represented priority or sensitive matters, and included public relations issues.

Factor 3 (Eigenvalue = 1.38) clustered those items that involved dealing with existing conflict among staff and/or situations requiring considerable interpersonal skill and insight in order to avoid staff

conflict or morale problems.

Factor 4 (Eigenvalue = 1.08) emphasized those items that required an understanding of the importance of organizational goal accomplishment and efficient methods of operation, along with a willingness to exercise management control in redirecting staff as needed to ensure positive organizational outcomes.

Two factor-scoring methods discussed by Gorsuch (1974, p. 238) were investigated. In method #1, items were allocated to the factor on which they loaded highest. In method #2, all items with salient loadings on a factor (twice the level required for significance) were allocated to the factor. For both methods, rounded weights in half-point intervals were used instead of exact loadings.

Table 18 gives the validity of total test factor scores (sum of individual factor scores) for each scoring method in predicting ratings by immediate and next-higher-level supervisors on the overall composite "same" level and "all" levels criterion measures. Table 19 provides estimates of the true validity of each method by correcting for unreliability in the criterion measures.

Table 18

OBTAINED VALIDITY COEFFICIENTS: CORRELATION OF SUM OF FACTOR SCORES FOR TWO FACTOR SCORING METHODS WITH OVERALL MEASURES OF PERFORMANCE BASED ON "SAME" AND "ALL" LEVELS RATINGS BY IMMEDIATE AND NEXT-HIGHER-LEVEL SUPERVISORS

	Immediate Supervisors		Next-Higher-Level Supervisors	
	Same Level Overall Measure	All Levels Overall Measure	Same Level Overall Measure	All Levels Overall Measure
Method 1	r = .32* (n = 272)	r = .32* (n = 275)	r = .34* (n = 238)	r = .34* (n = 239)
Method 2	r = .29* (n = 272)	r = .29* (n = 275)	r = .31* (n = 238)	r = .31* (n = 239)

* $p < .0001$

Table 19

BEST ESTIMATES OF TRUE VALIDITY OF SUM OF FACTOR SCORES
FOR TWO SCORING METHODS

	Immediate Supervisors		Next-Higher-Level Supervisors	
	Same Level Overall Measure	All Levels Overall Measure	Same Level Overall Measure	All Levels Overall Measure
Method 1	r = .41 (n = 272)	r = .43 (n = 275)	r = .44 (n = 238)	r = .45 (n = 239)
Method 2	r = .37 (n = 272)	r = .39 (n = 275)	r = .37 (n = 238)	r = .41 (n = 239)

ADDITIONAL VALIDITY DATA

The first study reported below was conducted by the U.S. Department of Army. All other studies reported in this section were conducted by the author.

Study #1

The U.S. Department of Army conducted a validation of the GMIB for 393 mid-level supervisors and managers (GS/13-GS/15) employed in the Civilian Personnel Officer job series (Mack and Lilienthal, 1991). Heterogeneous criterion measures were used, such as "quality of supervisory performance," "quantity of workgroup output," "positive work environment," "EEO/affirmative action," etc. Ratings on nine such criteria were summed to form a composite measure of overall performance. The reliability of the composite measure was .66.

Ratings on specific supervisory skills such as "leadership," "planning & organizing," etc., were not obtained. The GMIB was nevertheless found to have significant validity in predicting the heterogeneous composite measure of overall performance.

The authors found that the GMIB was valid at all levels of supervision included in the study, with no differences in validity x level. A test fairness study was conducted as well, and the authors concluded that the GMIB met test fairness requirements ("There were no differences in the magnitude of validity for blacks versus whites or males versus females... No significant differences were found in comparing black-white standard errors of estimates, slopes, and intercepts." p. 4).

The results of the study are shown in Table 20. This table reports estimated true coefficients (not all obtained coefficients were reported); some estimated true coefficients not reported in the article were sent to the author via personal communication in an internal briefing document prepared for the Civilian Personnel Administration Planning Board).

Table 20

STUDY #1: DEPARTMENT OF ARMY VALIDATION RESULTS FOR CIVILIAN PERSONNEL OFFICER SERIES, GRADES 13 - 15

Composite Measure of Overall Performance	Estimated True Validity	Sample Size
All candidates	.25	393
Blacks	.29	29
Whites	.26	329
First Level Supervisor	.23	299
Second Level & Above Supervisor	.22	94

Study #2

In a study of 191 candidates for first level supervisor, ratings were collected on the performance of the employees in their current positions. The candidates were employed across a variety of professional positions in a state government agency. Ratings of performance in "relation to others at the same organizational level" were obtained from either the subject's immediate or next-higher-level supervisor on the six performance dimensions. The raters also supplied ratings of the candidate's "potential to succeed" in supervision.

Ratings on 153 candidates by immediate supervisors were obtained; and on 38 candidates by next-higher-level supervisors (n=191). It should be noted that greater validity was found in predicting ratings by next-higher-level supervisors, consistent with a trend observed in the original GMIB validation study. This reflected significant differences in the views of the raters, and reliability for the limited available sample was lower than typically reported in the literature; thus, in order to avoid spuriously inflated estimated true validity coefficients, the reliability coefficients obtained in the original validation study were used (.62 for subjective overall rating, .61 for mechanical composite, and .51 for dimension ratings; and for the "combined" ratings, Spearman-Brown extended reliabilities of .77, .77, and .76, respectively).

It should be noted that this study represents a concurrent validation of the GMIB for professional staff positions in which the assessees were employed. However, it also represents a validation of the GMIB in predicting performance in skill areas known to be important to success in supervision and management.

Obtained and estimated true validities are given in Table 21. It should be noted that for the combined sample, all obtained validities were significant. Of the six validity coefficients computed for immediate and next-higher-level supervisors, five were significant.

Table 21

STUDY #2 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Rating: Immediate Supv.	.16*	153	.62	.20
Subjective Rating of Potential: Immediate Supv.	.19*	153	.62	.24
Mechanical Overall Rating: Immediate Supv. Same Level	.18*	153	.61	.23
Subjective Overall Rating: Next-Higher-Level Supv.	.34*	38	.62	.43
Subjective Rating of Potential: Next-Higher-Level Supv.	.28	38	.62	.36
Mechanical Overall Rating: Next-Higher-Level Supv Same Level	.34*	38	.61	.44
Subjective Overall Combined Rating	.20*	191	.77	.23
Subjective Overall Potential Combined Rating	.21*	191	.77	.24
Mechanical Overall Combined Rating: Same Level	.22*	191	.76	.25

* $p < .05$

Study #3

Table 22 gives the results of a study involving managers in a state government agency. The study utilized only the "same" level performance dimension ratings (in conjunction with the subjective overall performance rating).

Table 22

STUDY #3 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Rating: Immediate Supv.	.17	91	.56	.23
Mechanical Overall Rating: Immediate Supv. Same Level	.21*	91	.53	.29
Subjective Overall Rating: Next-Higher-Level Supv.	.28**	86	.56	.37
Mechanical Overall Rating: Next-Higher-Level Supv Same Level	.30**	86	.53	.41
Subjective Overall Combined Rating	.23*	80	.72	.27
Mechanical Overall Combined Rating: Same Level	.26**	80	.69	.31

* $p < .05$ ** $p < .01$

Study #4

In a study of 13 candidates for deputy chief of police in a major jurisdiction, two higher-level raters supplied ratings on: (1) overall performance relative to others at the "same" level; and (2) potential for success in higher management; and (3) combined ratings of the two raters on the six performance dimensions included in the original GMIB validation study (relative to "same" reference group). The obtained validity coefficients were .43, .48, and .43 ($p < .05$ for each). The Spearman-Brown reliabilities for the combined ratings on each variable were .79, .79, and .84, respectively; leading to estimated true validity coefficients of .48, .54 and .47.

Table 23

STUDY #4 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Rating: Immediate Supv.	.52*	13	.66	.64
Subjective Rating of Potential: Immediate Supv.	.56*	13	.65	.69
Mechanical Overall Rating: Immediate Supv. Same Level	.47*	13	.73	.55
Subjective Overall Rating: Next-Higher-Level Supv.	.25	13	.66	.31
Subjective Rating of Potential: Next-Higher-Level Supv.	.30	13	.65	.37
Mechanical Overall Rating: Next-Higher-Level Supv Same Level	.35	13	.73	.41
Subjective Overall Combined Rating	.43	13	.79	.48
Subjective Overall Potential Combined Rating	.48*	13	.79	.54
Mechanical Overall Combined Rating: Same Level	.43	13	.84	.47

* $p < .05$

Study #5

In a small-scale study of managers in a private sector firm, the managers were rank-ordered by the manager to which they all reported. The ranks were converted to T scores, which were then correlated with their respective scores on the GMIB. Since only one manager ranked the subjects, no measure of reliability was available and an estimated true validity coefficient is not reported. The results are given in Table 24.

Table 24

STUDY #5 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size
Ranking of Employees Based on Overall Performance	.80*	6

* $p < .05$

Study #6

In a study of candidates for promotion to the levels of sergeant, lieutenant, and captain in a state highway patrol agency, substantial validity coefficients were found but typically did not reach significance, more than likely due to the relatively small sample sizes. Results are shown for the sergeant and lieutenant positions and for the combined sample in Table 25. The captain sample, being less than five, is not separately shown. The results for subjective overall ratings of performance demonstrated low reliability, did not attain significance, and are not reported.

Table 25

STUDY #6 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Sergeant: Mechanical Overall Rating: Immediate Supv. Same Level	.15	23	.39	.24
Sergeant: Mechanical Overall Rating: Next-Higher-Level Supv. Same Level	.27	23	.39	.43
Sergeant: Mechanical Overall Combined Rating	.26	23	.56	.35
Lieutenant: Mechanical Overall Rating: Immediate Supv. Same Level	.37	15	.69	.45
Lieutenant: Mechanical Overall Rating: Next-Higher-Level Supv. Same Level	.24	15	.69	.29
Lieutenant: Mechanical Overall Combined Rating	.34	15	.82	.38
All Subjects: Mechanical Overall Rating: Immediate Supv. Same Level	.24	42	.38	.39
All Subjects: Mechanical Overall Rating: Next-Higher-Level Supv. Same Level	.29*	42	.38	.47
All Subjects: Mechanical Overall Combined Rating	.32*	42	.55	.43

* $p < .05$

Study #7

A total of 65 candidates for fire battalion commander and fire division chief went through the same assessment center, one part being the GMIB. The performance of the candidates was evaluated by two higher-level raters. Results are reported for the combined ratings in Table 26.

Table 26

STUDY #7 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Combined Rating	.34*	49	.78	.38
Mechanical Overall Combined Rating: "Same" Level	.35*	52	.77	.40

* $p < .005$

Study #8

In a study of 33 candidates applying for police lieutenant, immediate and next-higher-level raters produced completely different validity results; with the GMIB achieving positive or significant results for the immediate raters, but not the next-higher-level raters. Table 27 details the results of the study.

Study #9

Fifteen candidates for police lieutenant took the GMIB. All of the candidates were currently police sergeants in the same organization. The Chief of Police was very familiar with the performance of all of the candidates, and ranked them from highest to lowest in terms of suitability for promotion to the rank of lieutenant. The Chief's ranks were converted to T-scores which were then correlated with the scores obtained by the candidates on the GMIB. Since only one manager ranked the subjects, no measure of reliability was available and an estimated true validity coefficient is not reported. The results are given in Table 28.

Table 27

STUDY #8 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size	Reliability	True Validity
Subjective Overall Rating: Immediate Supv.	.53*	23	.60	.68
Subjective Overall Rating: Next-Higher-Level Supv.	- .06	30	.60	- - - -
Subjective Overall Rating of Potential: Immediate Supv.	.44**	23	.58	.58
Subjective Overall Rating of Potential: Next-Higher Supv.	- .03	31	.58	- - - -
Mechanical Overall Rating: Immediate Supv. Same Level	.31	22	.61	.40
Mechanical Overall Rating: Next-Higher Supv. Same Level	- .02	31	.61	- - - -

* $p < .005$ ** $p < .05$

Table 28

STUDY #9 VALIDATION RESULTS

Overall Criterion Measures	Obtained Validity	Sample Size
Ranking of Employees Based on Overall Suitability for Promotion	.67*	15

* $p < .005$

CORRELATIONS WITH OTHER INSTRUMENTS

Clients using the GMIB may also utilize one or more assessment exercises and/or conduct assessment centers. Table 29 summarizes data submitted by client organizations. The sample sizes, while not specified below, typically consist of small groups of candidates, usually 6 - 12. These results indicate a common pattern of substantial correlations with assessment exercises and assessment center results.

TABLE 29

GMIB CORRELATIONS WITH OTHER SELECTION DEVICES

Supervisory Level	Selection Process	Correlation
3rd Level Supervisor	2 Day A.C.	.55
2nd Level Supervisor	Dec. Making Sim. Interview	.84 .67
3rd Level Supervisor	1 Day A.C. Technical Knowledge	.47 .00
2nd Level Supervisor	LGD- Assigned LGD- Unassigned Oral Presentation	.61 .61 - .47
4th Level Supervisor	Interview	.31
1st Level Supervisor	Analysis/Report	.70
3rd Level Supervisor	2 Unassigned LGD's	.72
2nd Level Supervisor	2 Unassigned LGD's	.58
3rd Level Supervisor	2 Unassigned LGD's	.15
3rd Level Supervisor (2 Day A.C.)	Judgment Leadership Analysis Decisiveness Interpersonal	.90 .87 .79 .86 .71
4th Level Supervisor	2 Day A.C.	.91
2nd Level Supervisor	Interview, Watson-Glaser Plus 2 Simulation Tests	.76
3rd Level Supervisor	1 Unassigned LGD	.49
1st Level Supervisor	1 Unassigned LGD	.30
1st Thru 3rd Level Supervisors (2 Day A.C.)	Communication Skill Decision Making Management Style Leadership Interpersonal Relations Personal Skills Watson-Glaser Dimension Total	.41 .34 .28 .28 .14 .41 .36 .34

Test-Retest Results

Data on 173 candidates who have taken the GMIB more than once has been accumulated. These candidates took a parallel form of the GMIB on the second administration, which occurred between six months and two years after the first test. The mean on the first and second administrations = 17.0 and 19.6, respectively. The correlation for the two administrations = .74.

Racial and Sex Data

While the total GMIB data base has 18865 records at present, with a mean score = 18.83, racial and/or sex data is not available on all candidates in the data base. Table 30 provides results for those candidates for which data was provided by employers. It should be noted mean score differences across groups are substantially less than what is typically reported in the literature for ability and achievement tests (i.e., about 1 standard deviation). In addition, it should be noted that females score slightly higher than males.

Table 30

GMIB RACIAL AND SEX DATA

Group	Mean	S.D.	N
Racial Groups			
White	19.08	6.99	11718
Black	15.13	6.65	1721
Hispanic	16.98	7.18	956
Asian/Pacific Islander	16.77	7.15	416
American Indian/Alaskan Native	18.77	6.54	70
Filipino	17.74	7.23	99
Sex Groups			
Male	18.23	6.94	12664
Female	19.39	7.65	3389

Summary and Discussion

The GMIB is a new approach to in-basket testing. Items are scored individually based on explicit scoring guidance. The items are designed to test candidate skills in handling important, common management situations and are not tied to any particular type of organization or management position. The scoring guidance was developed based on the application of prevailing management theory and sound management principles to commonly occurring management situations and problems. The initial research study on the GMIB supported the appropriateness of the scoring guidance; but the results of

the research were also used to make refinements needed to achieve greater inter-rater reliability.

The GMIB can be scored in a highly reliable and efficient manner. The lowest obtained inter-rater reliability coefficient was .86. The simple mean of 42 obtained coefficients is .92, and if obtained coefficients are weighted by the number of in-baskets rated in the study, the mean is .93. Given such high reliability, a second rater does not significantly increase reliability; therefore, only one rater is required to score the GMIB. This makes the item-by-item scoring approach extremely attractive in comparison to traditional approaches.

The GMIB has proven to have substantial, significant validity across a series of criterion-related validity studies involving a variety of managers at differing organizational levels. The performance criteria that were included in the studies are routinely identified in job analysis studies as critical to success in management positions. The GMIB has consistently predicted composite ratings of these criteria, as well as subjective overall performance ratings.

The racial/sex data collected on the GMIB shows that the difference across all racial groups is less than .6 of a standard deviation. Women score slightly higher than men, but the difference is not of any practical significance.

The results of the factor analysis make it possible to profile candidates on their particular strengths and weaknesses. Traditional in-basket scoring approaches rate candidates on dimensions and attempt to achieve reliable profile information, although assessors frequently experience difficulties in clearly distinguishing between dimensions. Due to the item-by-item scoring approach of the GMIB, mathematically independent factor (dimension) scores can be readily generated for each candidate. This approach avoids the problems inherent in the traditional approach of attempting to make clear distinctions among dimensions which are often highly related and therefore not readily susceptible to such distinctions.

To date, there has only been one formal appeal challenging the job-relatedness/validity of the GMIB, and that appeal was denied in a formal administrative hearing after all evidence was heard. The GMIB has been used as a selection and development tool for both small and large organizations. It is the only narrative response format managerial skills test with a national data base.

REFERENCES

Brannick, Michael T., Michaels, Charles E., and Baker, David P. Construct validity of in-basket scores. Journal of Applied Psychology, 1989, 74, 957-963.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Gorsuch, Richard L. Factor Analysis. W. B. Saunders Company, Philadelphia, 1974.

Hinrichs, J. R. and Haanpera, S. Reliability of measurement in situational exercises: an assessment of the assessment center method. Personnel Psychology, l976, 29, 31-40.

Joines, Richard C. The item-by-item scored General Management In-Basket. Paper presented at the International Personnel Management Association Assessment Council annual conference, Philadelphia, Pennsylvania, 1987.

Joines, Richard C. The General Management In-Basket. Paper presented at the 17th International Congress on the Assessment Center Method, Pittsburgh, Pennsylvania, 1989.

Joines, Richard C. The General Management In-Basket. Paper presented at the 19th International Congress on the Assessment Center Method, Toronto, Canada, 1991.

Kesselman, G. A., Lopez, F. M., and Lopez, F. E. The development and validation of a self-report scored in-basket test in an assessment center setting. Public Personnel Management Journal, l982, 11, 228-238.

Kim, J. Factor analysis. In Nie, N.J., Hull, C.H., Jenkins, J.G., Steinbrenner, K., Brent, D.H. (Eds.), SPSS: Statistical Package for the Social Sciences, pp. 468-514. McGraw-Hill, New York, 1975.

Kraus, J. C. The multiple-choice in-basket exercise as developed and used by the N. J. Department of Civil Service. Paper presented at the International Personnel Management Association Assessment Council annual conference, San Francisco, California, 1986.

Lopez, F. M. Evaluating executive decision-making: The in-basket technique. American Management Association, Inc. (AMA Research Study 75), l966.

Mack, M.J. and Lilienthal, R.A. Implementation and evaluation of an in-basket test for supervisory referral, Military Testing Association 33rd Annual Conference, 1991.

Sackett, P.R., & Dreher, G.F. Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 1982, 67, 401-410.

Sackett, P.R., & Dreher, G.F. Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. Journal of Applied Psychology, 1984, 69, 187-190.

Schippmann, J.S., Hughes, G.L. and Prien, E.P. The use of structured multi-domain job analysis for the construction of assessment center methods and procedures. Journal of Business and Psychology, 1987, 1, 353-366.

Schippmann, J. S., Prien, E. P., and Katz, J. A. Reliability and validity of in-basket performance measures. Personnel Psychology, 1990, 43, 837-859.

Thornton, G. C. III and Byham, W. C. Assessment Centers and Managerial Performance. Academic Press, New York, l982.

Wollowick, H. B. and McNamara, W. J. Relationship of components of an assessment center to management success. Journal of Applied Psychology, l969, 53, 348-352.