AN EXAMINATION OF SEVERAL IN-BASKET SCORING STRATEGIES AND THEIR EFFECT ON RELIABILITY AND CRITERION-RELATED VALIDITY

by

KAREN P. HARLOS

B. A. University of British Columbia, 1982

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS in

THE FACULTY OF GRADUATE STUDIES

Department of Psychology

We accept this thesis as conforming to the required standing

THE UNIVERSITY OF BRITISH COLUMBIA

November, 1992

© Karen P. Harlos, 1992
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Psychology
The University of British Columbia
Vancouver, Canada

Date February 5, 1993
ABSTRACT

The in-basket exercise, a paper-and-pencil measure of administrative ability, is an assessment technique characterized by complex, often subjective scoring procedures which have limited wide-scale applications of this popular instrument. Because the literature affords no clear classification system for this and related exercises, a structure which clarified the definitions of and relationships among management games, management simulations, work sample tests, and in-basket exercises was introduced. The primary purpose of the present research was to investigate the cross-sample generalizability of several strategies for scoring the in-basket exercise, including a reduced-item scoring approach in which an optimal subset of items was identified for scoring. The preliminary studies upon which the present research is based are also discussed in the present work. In addition to examining the impact of these scoring strategies on reliability and criterion-related validity, consideration was given to addressing long-standing concerns of in-basket training- and scoring time demands.

Three hundred and twenty-one entry-level employees from a large western Canadian utility company were administered the same in-basket exercise previously applied in a different Canadian utility company. Contrary to expectations, the shrinkage in validity using an empirically-based scoring key was substantial, pointing to the selection of a more logically-derived panel key as the method of choice. The introduction of a new cognitive-based measure of in-basket performance showed promising results. In addition, the reduced-item scoring approach did not result in significant losses in reliability or criterion-related validity, thus allowing substantial reductions in training- and scoring-time. Implications for in-basket development are also considered.
TABLE OF CONTENTS

ABSTRACT ................................................................. ii

LIST OF TABLES ............................................................ x

LIST OF FIGURES .......................................................... xiii

ACKNOWLEDGMENT ........................................................ xiv

INTRODUCTION
The In-Basket Exercise ....................................................... 1
  Definition and Design .................................................... 1
The In-Basket as a Work Sample Test ...................................... 2
  Legal influences on work sample testing ............................. 2
  Economic influences on work sample testing ....................... 3
Classification of Scoring Methods ........................................... 4
  The objective method .................................................. 5
  The subjective method ................................................ 6
  The combination method .............................................. 7
Rationale for the Present Study ............................................ 7
  Need for Further In-Basket Research ................................. 7
Overview of Preliminary Studies ........................................... 9
  Study 1 ............................................................... 9
  Study 2 ............................................................. 10
The Present Study ......................................................... 11

LITERATURE REVIEW
Section 1: Management Simulations ....................................... 13
  History and Development ............................................. 13
  Terms and Definitions ............................................... 14
  Historical Background of Management Simulations ............... 16
Management Simulations as Work Sample Tests ......................... 19
Issues in Management Simulation Design and Research ............... 27
  Key Theoretical Issues: Realism and Fidelity ...................... 27
    Realism ......................................................... 27
### TABLE OF CONTENTS (cont.)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity</td>
<td>28</td>
</tr>
<tr>
<td>Validity in the Context of Management Simulations</td>
<td>31</td>
</tr>
<tr>
<td>Content validity</td>
<td>32</td>
</tr>
<tr>
<td>Construct validity</td>
<td>32</td>
</tr>
<tr>
<td>Criterion-related validity</td>
<td>34</td>
</tr>
<tr>
<td>Summary of Section 1: Management Simulations</td>
<td>36</td>
</tr>
<tr>
<td>Section 2: Psychometric Properties of the In-Basket Exercise</td>
<td>37</td>
</tr>
<tr>
<td>The Seminal In-Basket Work</td>
<td>37</td>
</tr>
<tr>
<td>Goal of Frederiksen's work</td>
<td>37</td>
</tr>
<tr>
<td>Identification of Performance Dimensions</td>
<td>38</td>
</tr>
<tr>
<td>Development and selection of exercise problems</td>
<td>38</td>
</tr>
<tr>
<td>Development of the scoring method</td>
<td>38</td>
</tr>
<tr>
<td>Significance</td>
<td>40</td>
</tr>
<tr>
<td>Introductory Comments on In-Basket Empirical Findings</td>
<td>41</td>
</tr>
<tr>
<td>Reliability of the In-Basket Exercise</td>
<td>46</td>
</tr>
<tr>
<td>Alternate Form Reliability Estimation</td>
<td>46</td>
</tr>
<tr>
<td>Conceptual foundation</td>
<td>46</td>
</tr>
<tr>
<td>Empirical findings</td>
<td>47</td>
</tr>
<tr>
<td>Summary</td>
<td>48</td>
</tr>
<tr>
<td>Internal Consistency: Split-Half and Alpha Coefficients</td>
<td>48</td>
</tr>
<tr>
<td>Conceptual foundation</td>
<td>48</td>
</tr>
<tr>
<td>Empirical findings</td>
<td>49</td>
</tr>
<tr>
<td>Summary</td>
<td>51</td>
</tr>
<tr>
<td>Inter-Rater Reliability Estimation</td>
<td>52</td>
</tr>
<tr>
<td>Conceptual foundation</td>
<td>52</td>
</tr>
<tr>
<td>Empirical findings</td>
<td>52</td>
</tr>
<tr>
<td>Summary</td>
<td>52</td>
</tr>
<tr>
<td>General Summary of In-Basket Reliability Results</td>
<td>53</td>
</tr>
<tr>
<td>Validity of the In-Basket Exercise</td>
<td>54</td>
</tr>
<tr>
<td>Criterion-related Validity</td>
<td>55</td>
</tr>
<tr>
<td>Air Force in-baskets</td>
<td>55</td>
</tr>
<tr>
<td>School administration in-baskets</td>
<td>55</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS (cont.)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Port of New York Authority studies</td>
<td>60</td>
</tr>
<tr>
<td>The Management Progress Study</td>
<td>64</td>
</tr>
<tr>
<td>Cross' work with the school administration in-basket</td>
<td>66</td>
</tr>
<tr>
<td>IBM study</td>
<td>67</td>
</tr>
<tr>
<td>The General Electric research</td>
<td>68</td>
</tr>
<tr>
<td>A &quot;leadership&quot; in-basket exercise</td>
<td>71</td>
</tr>
<tr>
<td>A return to the self-report scored in-basket exercise</td>
<td>74</td>
</tr>
<tr>
<td>A predictive validation study</td>
<td>77</td>
</tr>
<tr>
<td>A quickly-scored in-basket</td>
<td>78</td>
</tr>
<tr>
<td>Summary of criterion-related validity evidence</td>
<td>78</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>79</td>
</tr>
<tr>
<td>School administration in-baskets</td>
<td>79</td>
</tr>
<tr>
<td>Frederiksen (1966)</td>
<td>80</td>
</tr>
<tr>
<td>A parallel form study</td>
<td>81</td>
</tr>
<tr>
<td>The measurement of participative decision-making</td>
<td>83</td>
</tr>
<tr>
<td>Another multitrait-multimethod study</td>
<td>84</td>
</tr>
<tr>
<td>Factor-analytic studies</td>
<td>85</td>
</tr>
<tr>
<td>Summary of construct-related validity evidence</td>
<td>86</td>
</tr>
<tr>
<td>Content Validity</td>
<td>86</td>
</tr>
<tr>
<td>General Summary of In-Basket Validity Results</td>
<td>87</td>
</tr>
<tr>
<td>Summary of Section 2: Psychometric Properties of the In-Basket</td>
<td>87</td>
</tr>
<tr>
<td>Exercise</td>
<td>89</td>
</tr>
<tr>
<td>Section 3: A Review of Current In-Basket Scoring Strategies</td>
<td>90</td>
</tr>
<tr>
<td>Limitations of a Literature-Based Review</td>
<td>90</td>
</tr>
<tr>
<td>Review of Current In-Basket Scoring Methodology in Industry</td>
<td>92</td>
</tr>
<tr>
<td>The Multiple-Choice In-Basket Exercise</td>
<td>93</td>
</tr>
<tr>
<td>General Management In-Basket Exercise</td>
<td>94</td>
</tr>
<tr>
<td>ETS Consolidated Fund In-Basket Exercise</td>
<td>97</td>
</tr>
<tr>
<td>An example of the combination scoring method</td>
<td>98</td>
</tr>
</tbody>
</table>
Summary of current in-basket scoring methodology in industry ........................................... 100

Summary of Section 3: A Review of Current In-Basket Scoring Strategies ........................................... 101

PRELIMINARY STUDIES

Study 1 .................................................................................................................. 102
Method .................................................................................................................. 102
Participants and Setting ....................................................................................... 102
Materials .............................................................................................................. 102
The Telephone Supervisor In-Basket Exercise ................................................... 102
Criterion measurement ....................................................................................... 103
Procedure ............................................................................................................ 104
Development of the TSIB .................................................................................... 104
Identification of the TSIB dimensions ............................................................... 105
Development of the TSIB dimension scoring keys .......................................... 107
General scoring protocol ................................................................................... 108
Results ................................................................................................................. 112
Reliability ............................................................................................................ 112
Criterion measurement ....................................................................................... 112
TSIB dimensions ................................................................................................ 112
Validity ................................................................................................................ 114
Discussion .......................................................................................................... 114

Study 2 ................................................................................................................. 116
Method ................................................................................................................. 116
Participants and Setting ....................................................................................... 116
Procedure ............................................................................................................. 117
Modification of the Quality of Judgement scoring key ...................................... 117
A new, expanded panel ...................................................................................... 117
Derivation of the new scoring keys for Quality of Judgement .......................... 118
Derivation of the reduced-item sets .................................................................. 122
### TABLE OF CONTENTS (cont.)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>122</td>
</tr>
<tr>
<td>Reliability</td>
<td>122</td>
</tr>
<tr>
<td>Validity</td>
<td>123</td>
</tr>
<tr>
<td>Intercorrelations</td>
<td>123</td>
</tr>
<tr>
<td>Construct validity</td>
<td>128</td>
</tr>
<tr>
<td>Discussion</td>
<td>129</td>
</tr>
<tr>
<td>RATIONALE AND HYPOTHESES FOR THE PRESENT STUDY</td>
<td>133</td>
</tr>
<tr>
<td>General Summary</td>
<td>133</td>
</tr>
<tr>
<td>Proposal and Hypotheses</td>
<td>135</td>
</tr>
<tr>
<td>Criterion measurement</td>
<td>136</td>
</tr>
<tr>
<td>Quality of Judgement Scores</td>
<td>136</td>
</tr>
<tr>
<td>Predictive validities of scores based on the</td>
<td></td>
</tr>
<tr>
<td>Panel, Empirical and Merged scoring keys</td>
<td></td>
</tr>
<tr>
<td>Additional analyses based on the</td>
<td></td>
</tr>
<tr>
<td>Panel scoring key</td>
<td>137</td>
</tr>
<tr>
<td>Additional Performance Dimensions in the TSIB</td>
<td>140</td>
</tr>
<tr>
<td>Understanding of Situation Questionnaire</td>
<td>140</td>
</tr>
<tr>
<td>Productivity</td>
<td>142</td>
</tr>
<tr>
<td>In-Basket Stylistic Dimensions</td>
<td>143</td>
</tr>
<tr>
<td>Additional New Measures</td>
<td>144</td>
</tr>
<tr>
<td>Number of Priority Items attempted</td>
<td>144</td>
</tr>
<tr>
<td>Scorer's Impression of Involvement</td>
<td>145</td>
</tr>
<tr>
<td>METHOD</td>
<td></td>
</tr>
<tr>
<td>Participants and Setting</td>
<td>146</td>
</tr>
<tr>
<td>Assessment Measures</td>
<td>147</td>
</tr>
<tr>
<td>Predictor measures</td>
<td>147</td>
</tr>
<tr>
<td>Criterion measures</td>
<td>147</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS (cont.)

RESULTS

Overview of Study Design and Data Analysis.........................151
  Design..................................................................151
  Data Analysis.........................................................151
    Reliability .........................................................151
    Criterion-related validity ....................................152
    A brief comment on the Results ..............................153

Major Reliability and Criterion-related Validity Findings ........153

Reliability of the Performance Appraisal Dimension
  Scores ..................................................................153

Cross-Validation of the Quality of Judgement Scores ............154

Other Performance Dimensions in the TSIB .......................157
  Understanding of Situation Questionnaire .....................157
  Productivity .......................................................162

TSIB Stylistic Dimensions ........................................163

Reliability of the TSIB dimensions ................................170

Additional New Measures ...........................................170
  Number of Priority Items Attempted ............................170
  Scorer's Impression of Involvement ..............................170

Factor Analysis of the TSIB Action Elements .....................174

A Series of Factorings ..............................................174
  The first factor analysis .......................................174
  The second factoring ............................................175
  The third factoring ..............................................176

Interpretation of the factors ......................................176
  Factor I: Consultation and Discussion .........................176
  Factor II: Independent Decision-Making .......................177
  Factor III: Company-Based Decision-Making .................178
  Factor IV: Supervising Staff ...................................178
  Factor V: Integration and Analysis ............................179

Assignment of Factors to Action Elements ........................179
# TABLE OF CONTENTS (cont.)

## DISCUSSION
- Reliability and Criterion-related Validities of Selected Dimensions of the TSIB .......................................................... 185
- Cross-Validation of the Quality of Judgement Scores .......... 185
- Predictive validities of scores based on the Panel, Empirical, and Merged keys ................................................. 185
- Additional analyses based on the Panel scoring key ......................................................................................... 189
- Additional Performance Dimensions in the TSIB ................. 196
- Understanding of Situation Questionnaire ...................... 196
- Productivity ........................................................................ 198
- TSIB Stylistic Dimensions .................................................... 201
- Reliability of the TSIB dimensions .................................. 202
- Additional New Measures .................................................... 204
  - Scorer's Impression of Involvement .............................. 204
  - Number of Priority Items Attempted ............................. 204
- Selection of the Optimal Item Subset: Integrating Empirical Results and Practical Issues ................................. 205
- Factor Analysis of the TSIB Action Elements ................ 206
  - A Note Regarding A Linear Combination of the TSIB Dimensions .......................................................... 209
- Summary and Conclusions ............................................... 210

## REFERENCES ........................................................................ 214

## APPENDICES
- Appendix A List of Selected Cognitive Ability and Personality Measures Used in the Construct Validity Analysis .......... 223
Table 1 Review of Major Empirical Findings on the In-Basket .................. 43
Table 2 Inter-Rater Reliability Estimates for the TSIB ............................. 113
Table 3 Validity Coefficients between Overall Performance Criterion (OPC) 
and TSIB Dimension Scores .......................................................... 115
Table 4 Alpha and Stepped-up Split-Half Reliability Estimates for Quality of 
Judgement Scores Obtained from Three Items Sets as Scored by the 
Merged Key .................................................................................... 124
Table 5 Bivariate Correlations between Overall Performance Criterion and 
Quality of Judgement Scores Obtained from Four Scoring Keys and 
from Three Item Sets ................................................................. 125
Table 6 Intercorrelations of Dimension Scores based on the 21-item Set: 
Stylistic Dimensions Scored by Logical Keying System and Quality of 
Judgement Scored by Panel 1 (Liberal) Key ................................. 126
Table 7 Intercorrelations of Dimension Scores based on the 8-item Set: 
Stylistic Dimensions Scored by Logical Keying System and Quality of 
Judgement Scored by Panel 1 (Liberal) Key ..................................... 127
Table 8 Reliability Analysis (Alpha Coefficient) of the Five Performance 
Appraisal Dimensions ................................................................ 155
Table 9 Bivariate Correlations between Overall Management Performance 
and Quality of Judgement Scores Obtained from Three Scoring Keys 
and from Three Item Sets .............................................................. 156
Table 10 Bivariate Correlations between Overall Management Performance 
and Quality of Judgement Scores Obtained Using Scoring Keys based 
on Different Methods of Combining Panel Judgements (Liberal or 
Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), 
and Applied to Different Datasets (Company 1 or Company 2) ....... 158
# LIST OF TABLES (cont.)

| Table 11 | Validity Coefficients between Overall Management Performance and Understanding of Situation Questionnaire Total Score (32-item); Alpha Reliability Estimates for Total Score | 160 |
| Table 12 | Validity Coefficients between Overall Management Performance and Understanding of Situation Questionnaire Total Score (20-item) and Reliability Estimates for Total Score | 161 |
| Table 13 | Validity Coefficients between Overall Management Performance and Productivity as Measured by Several Individual Units of Output, Linear Composites, and Standard Total Action Indices | 164 |
| Table 14 | Validity Coefficients between Overall Management Performance and TSIB Stylistic Dimension Scores | 165 |
| Table 15 | Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores Based on the 8-Item Set | 167 |
| Table 16 | Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores Based on the 10-Item Set | 168 |
| Table 17 | Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores Based on the 12-Item Set | 169 |
| Table 18 | Reliability Estimates of the TSIB Dimensions | 171 |
| Table 19 | Bivariate Correlations between Number of High Priority Items Attempted and Overall Management Performance (OMP) | 173 |
| Table 20 | Primary-Factor Intercorrelations | 180 |
| Table 21 | Factor-Based Scale Score Intercorrelations | 182 |
| Table 22 | Validity Coefficients between Overall Management Performance (OMP) and TSIB Factor-Based Scale Scores and Reliability Estimates | 183 |
**LIST OF TABLES (cont.)**

Table 23  Summary of Validity Coefficient between Overall Management Performance and Quality of Judgement Scores Obtained Using Scoring Keys based on Different Methods of Combining Panel Judgements (Liberal or Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), and Whether Panel Origin and Dataset Match (Intra- or Inter-company analysis) .............................................................. 192

Table 24  Mean Validity Coefficients Determined by Aggregation Across Method of Combining Panel Judgements, Origin of Panel, and Nature of Analysis (Intra- or Inter-company). .................................................................................. 194
LIST OF FIGURES

Figure 1  Relationship between simulations and work sample tests .................. 22

Figure 2  Relationships among simulations, work sample tests,
management simulations, management games, and
in-basket exercises .......................................................... 26

Figure 3  Sample Item from the Telephone Supervisor In-Basket Exercise .......... 110

Figure 4  Excerpt from the Scoring Manual for Sample Item
(shown in Figure 3) .................................................................. 111

Figure 5  Design grid for planned panel analyses in terms of origin
of the Panel key, of method of combining panel judgements,
and of the dataset to which the keys will be applied ...................... 138

Figure 6  Sample Item from the Understanding of Situation Questionnaire ........ 141
ACKNOWLEDGEMENT

I would like to thank my committee members, Dr. Ralph Hakstian, Dr. Jerry Wiggins, and Dr. Beth Haverkamp for their contributions toward the organization and preparation of what became a lengthy tome. As my advisor, Dr. Hakstian deserves particular thanks for his assistance in the provision of research resources and his guidance in the direction and evaluation throughout both the preliminary and present studies.

In addition, Lois Crooks and Peggy Mahoney of ETS, who provided me with patient and pleasurable instruction in the scoring of in-baskets, are gratefully acknowledged. In the present study, further assistance in scoring and administration was provided by Derek Sam, Susan Greaves, and Rodney Hicks. Several employees of the Human Resources Department of the second company were very important to the successful completion of the present cross-validation: Cherie Chapman, Paula Weber, Carol Harper, and Joy Powell.

Lastly, I must thank Danny for his cherished intellectual and emotional support.
INTRODUCTION

The identification and development of future administrators is of great interest to company executives and industrial psychologists. Psychometric assessment of administrative ability is most commonly conducted by means of the in-basket exercise, which attempts to simulate the job conditions of managers by providing typical samples of managerial work and by requiring the candidate to take action on the problems presented. The in-basket exercise was first developed in the late 1950's as a training tool for officers in the U.S. Air Force (Frederiksen, Saunders & Wand, 1957). Its applications have now expanded to include certification and assessment for selection and promotion decisions. The in-basket exercise is an established feature of the managerial assessment centre, which is a method (rather than location) characterized by multiple assessment procedures and pooled judgements of multiple assessors to evaluate the performance of managerial candidates (Bray & Grant, 1966; Cascio, 1987). In fact, a survey of large American companies with assessment centres found that 31 out of 34 used the in-basket; it was also rated the most popular exercise among job candidates (Bender, 1973). More recently, Gaugler, Bentson and Pohley (cited in Thornton, 1992) surveyed the types of exercises used in over 200 assessment centres and found that 81% regularly used the in-basket.

The In-Basket Exercise

Definition and Design

The in-basket exercise consists of a stack of letters, memos, reports and related materials that are presumed to have accumulated in the in-basket of a manager. The participant is directed to deal with the materials as if he or she were actually on the job as the manager. Performance on the exercise can be scored on various dimensions that
are important to the successful performance of managerial work. Although usually group-administered, the in-basket exercise allows the participant to act independently by reading and responding in writing to the mountain of memos, letters and other documents. Typically, a job analysis is conducted to ensure that the problems and situations presented in the in-basket exercise are representative and accurate samples of work and so will be valid measures of on-the-job performance.

Generally, the in-basket exercise is composed of two substantive parts and a procedural element that have become the traditional model for design and construction of this instrument (Lopez, 1966). The first part is made up of background materials such as instructions, an organizational chart, and job descriptions to help orient participants to the fictitious company and outline the expectations and guidelines for responding. The second part involves the actual set of items or problems (letters, memos, etc.) to which the participant responds, as well as the materials needed to respond (letterhead, memo forms, paper clips, etc.). The procedural element of the in-basket exercise refers to the method by which the participant's actions are evaluated, whether scored objectively from written responses, or assessed more subjectively through an interview with a trained evaluator, for example.

The In-Basket as a Work Sample Test

As noted, the problems presented in the in-basket exercise are based on samples of managerial work. Accordingly, the in-basket can be considered as a member of a broader category of instruments known as "work sample" tests, which present samples of work representative of the target job and measure candidates' ability to do the work presented in the test.

Legal influences on work sample testing. In recent years, work sample testing has undergone extensive growth in response to legal pressures regarding the fairness of
selection procedures, especially in relation to women and minorities. In Griggs v. Duke Power Company (1971), the U. S. Supreme Court issued a landmark ruling which established several key principles in employment discrimination cases, including the requirement that selection tests be job relevant (that a given requirement for employment is related to job performance). The importance of job relevance for selection measures was later reaffirmed by the U.S. Supreme Court in Albemarle Paper Co. v. Moody (1975). In order to comply with the judicial requirement of fairness in selection procedures, the U. S. federal government published the Uniform Guidelines on Employee Selection Procedures (1978), which articulates three validation procedures to establish the legal defensibility of selection measures, namely construct-, criterion-, and content-validation (these will be discussed in greater detail in Section 1 of the Literature Review).

By their design, work sample tests are likely to satisfy the guideline of content validity (i.e., the extent to which the exercise truly measures the skills and abilities required for and used in the target job). Because work sample tests are based on representative samples of work from the target job, they are designed to be content valid, and therefore, are job relevant. As a work sample test, the development of the in-basket exercise is also based on the notion of content validity and job relevance. Thus, the legal requirement of job relevance and the federal guideline of content validity have significantly influenced the growth and development of work sample testing and in-basket exercises.

Economic influences on work sample testing. In addition to legal pressures to ensure fairness, selection procedures have also been subjected to economic pressures resulting from the current recession. Limited human resources budgets have called for assessment measures which can be administered and scored efficiently, thereby reducing costs. Thus, assessment measures which are psychometrically sound (reliable
and valid), legally defensible (job relevant and fair), as well as economically viable (quickly-administered and quickly-scored) are required.

While work sample tests tend to be psychometrically sound and legally defensible, their economic viability may be limited. Hunter and Hunter (1984) pointed out that work sample tests usually incur high costs for the initial development of the instrument, for further re-development if the target job changes, and for test administration. Furthermore, it will be seen shortly that, because of the complexity of in-basket scoring, it is not unusual for an in-basket allowing free-format or open-ended responses to require from 1 1/2 to 5 hours to evaluate, whereas many established personality inventories used for managerial assessment can be fully scored in under five minutes. Consequently, the use of the in-basket as a measure of administrative ability is a costly component of managerial assessment, given its high development and administration costs, as well as the considerable time required both to train scorers and for trained scorers to evaluate a completed in-basket. A unique challenge, then, facing in-basket exercise developers has been in meeting the concerns of company executives to improve economic efficiency in an increasingly competitive corporate market.

Classification of Scoring Methods

Lopez (1966) noted that, despite the apparent utility of this instrument, "the major weakness lay in its complex, tedious scoring process" (p. 108). In an effort to simulate real-world conditions, the typical in-basket exercise allows the candidate to respond in a written free-format fashion. Given a free-format design, subjectivity and lengthy time demands have, to varying degrees, affected all methods of in-basket scoring. Generally, the training process required to reliably score in-baskets is complex, given the number and scope of the problems presented and the judgement often needed to determine candidates' intent. Even when scorers are fully trained, it is
a complex process for them to decipher and interpret written responses, which are often hastily scrawled as candidates try to complete as much as work they can in the time available.

There are three main methods of scoring in-basket performance: (a) objective (psychometric), (b) subjective (interpretive), and (c) a combination of the two. At this time, no known information exists regarding the relative rates of use of these in-basket scoring approaches (B. Gaugler, personal communication, June 30, 1992; G. Thornton, personal communication, June 30, 1992).

The objective method. The "traditional objective" scoring approach which follows principles established by Frederiksen et. al. (1957) and later refined by Lopez (1966) focuses primarily on candidates' free-format written responses. In-basket scoring using this method attempts to break down these open-ended, free-format responses into molecular units of action (called action elements) which are then quantitatively scored for style and content. This psychometrically-based objective method can often require 1 1/2 to 2 hours per exercise to score, which, for large-scale applications, can present prohibitive costs in terms of both time (i.e., time required for training scorers and scoring time itself) and money (i.e., administration costs).

In response to these concerns of lengthy scoring time and high costs, a recent development in in-basket design has been away from the free-format response mode and toward a multiple-choice, self-report response mode (Kesselman, Lopez & Lopez, 1982; Morris, 1991). Candidates still may initially respond to items in written, open-ended form, but, upon completion of the exercise, this development then requires candidates to indicate the course of actions they took using a self-report multiple-choice form. Only this self-report form is then evaluated by scorers; the open-ended written responses are typically not examined by scorers. This design change has resulted in a
modification to the traditional objective scoring method because the scoring of only the multiple choice responses precludes the scoring of action elements derived from open-ended responses. A scoring key is applied to the candidate's choice of the response alternatives provided, thereby eliminating scorers' interpretive judgements in response assessment.

Although objectivity and efficiency in scoring are greatly improved by restricting responses, a clear shortcoming of the self-report response mode is the new threat of motivational distortion. Because scorers rarely consider the open-ended responses, participants may, in hindsight, select a self-report course of action that they believe is superior (and is contrary to the course of action indicated in their open-ended responses). Consequently, deception is a clear possibility for those in-basket exercises combining open-ended responses with a self-report, multiple-choice response format. In addition, concerns of lack of realism and excessive restriction of candidates' responses make multiple-choice in-basket exercises a controversial development in in-basket design and scoring.

The subjective method. In this commonly-used approach, the primary source of information about in-basket performance comes from an interview with the candidate upon completion of the in-basket (Bray & Grant, 1966; Frederiksen, Jensen & Beaton, 1972). The scorer's interpretive judgements of the candidate's verbal (and possibly written) responses then form the basis of a report describing administrative performance. A consequence of the highly subjective, interpretive nature of this scoring method is its inherent unreliability. Furthermore, given this unreliability, empirical evidence of a predictive relationship between in-basket "scores" and performance criteria (i.e., criterion-related validity) would be difficult to provide using the subjective scoring method. Like the objective method, the subjective scoring method also involves similarly significant training- and scoring-time demands.
The combination method. The combination method of in-basket scoring typically involves both an interview with the candidate and an objective examination of the candidate's written responses. This approach, which continues to be widely-used in managerial assessment centres, requires even greater time and financial demands than the objective scoring method. Even with a trained assessor, it is not unusual for in-basket scoring using the combination method to require 2 to 5 hours per exercise when used in managerial assessment centres (Thornton & Byham, 1982).

Rationale for the Present Study

Need for Further In-Basket Research

Despite its forty-year history and widespread use, relatively few empirical studies examining the reliability and validity of the in-basket exercise have been conducted, perhaps because its high face validity has lead to an over-reliance on the in-basket's appearance of job-relatedness (Gill, 1979). Gill's observation, although made over a decade ago, remains valid today; Schippmann, Prien and Katz (1990), in a review of the psychometric properties of in-basket measures concluded, "Whatever the reason, research and opinion about the in-basket can best be described as incomplete and arrested in time" (p. 856).

In addition, as will be shown in the Literature Review, the quality of the research that is reported in this area can be criticized for its lack of concrete information on methods, procedures, and results. Moreover, great variations in in-basket construction, performance dimensions, and external non-instrument criterion measures (e.g., salary progression, rate of promotion, supervisory ratings) make meaningful comparisons across studies difficult. There is continued controversy as to whether the psychometric evidence from published studies is strong enough to warrant
the in-basket's extensive and firm foundation in assessment programs (Gill, 1979; Schippmann et. al., 1990).

As noted earlier, assessment measures which are psychometrically sound (reliable and valid), legally defensible (job relevant and fair), as well as economically viable (quickly-administered and quickly-scored) are required. Given their classification as work sample tests, in-basket exercises are likely to be legally defensible. At the same time, the psychometric evidence from published studies is inconclusive regarding the in-basket's extensive and firm foundation in assessment programs. Moreover, the economic viability of the in-basket exercise has been severely limited because of high development, administration, and scoring costs (resulting from extensive training of scorers and scoring time itself). Excessive time demands stemming from the complex nature of in-basket scoring have limited the wide-scale application of the exercise with large numbers of candidates.

However, if modifications to the scoring system could be made which significantly reduced the time demands but did not significantly decrease the validity and reliability of the instrument, perhaps the in-basket exercise could enjoy an expanded application with large numbers of candidates for training and selection. As noted earlier, the scoring of the in-basket has long been recognized as the weak link in this important and popular instrument. Nearly forty years after the initial development of the in-basket exercise, Brannick, Michaels, and Baker (1989) cited comparisons of scoring systems in in-basket research as a key area for much-needed investigation.

Accordingly, it is this aspect of the in-basket, the scoring method, which is the focus of the present study. More specifically, because of the inherent unreliability of the subjective and, to a lesser extent, combination methods of scoring, modifications to the traditional objective scoring method will be examined. This will provide the
opportunity to assess the criterion-related validities of the modifications to the traditional objective scoring method within a free-format response mode in-basket exercise.

The present study is the result of efforts to build on previous work by Hakstian, Woolsey and Schroeder (1986) which outlined an attempt to develop an in-basket exercise with an objective scoring key and modest training- and scoring-time demands. Two preliminary studies (described below under Study 1 and Study 2) were conducted which made modifications to both the instrument itself and scoring strategies to further the goals of Hakstian et al. (1986). A more complete discussion of the two studies is found in the Preliminary Studies section of the present work.

The present study is a cross-validation of the findings of Study 2 which examined the impact of several scoring strategies on the psychometric properties of the instrument. In addition, using a novel approach of scoring an optimal subset, rather than the entirety of items in the exercise, industry concerns regarding time and cost required for training and actual scoring of the in-basket are considered in the evaluation of the instrument.

Overview of Preliminary Studies

Study 1. Nearly three years ago, a sample of 321 entry-level managers from a large Canadian utility company were administered a newly-revised in-basket exercise as part of a concurrent validity study for a managerial assessment battery. The psychometric properties of the instrument were determined; inter-rater reliability and criterion-related validity of seven new performance dimensions were assessed. In addition, in order to reduce scoring time, a smaller, optimal subset of items was identified so that this smaller subset (rather than the entire set of items) would be scored. One of the seven in-basket performance dimensions was selected as the most
critical aspect of administrative performance, and so became the primary predictor used in this and subsequent studies. The derivation of the optimal item-subset was made on the basis of results from this most critical in-basket performance dimension as measured by one scoring key. The Preliminary Studies section contains an outline of the development of the in-basket used in this (and the present) study and the details of the scoring keys generated for the seven performance dimensions, and also contains a discussion of the procedure used to derive the item-subset. The empirical results are summarized and briefly discussed under Study 1 in the Preliminary Studies section.

**Study 2.** Analysis of the findings from Study 1 led to a further development designed to improve the criterion-related validity of the in-basket exercise and preserve the use of the optimal subset of items for reduced scoring time. To improve the validity for the critical performance dimension chosen in Study 1, the scoring key for this dimension was modified. Specifically, three separate scoring strategies were developed and applied to this most critical in-basket performance dimension. In addition, the item composition of the optimal item-subset identified in Study 1 was also evaluated using the new scoring strategies. Actually, in addition to the original subset of items (identified in Study 1), two new subsets of items were also scored (for the critical performance dimension) using the three scoring strategies in order to select the most predictive scoring method as applied to the most optimal subset of items. In this way, the effects of both the item-subset composition and the scoring method could be examined at the same time. The findings suggested that an empirically-based scoring method yielded the highest criterion-related validity coefficients (which were substantially higher than the validity using the Study 1 scoring method), and one particular subset emerged as the most optimal (psychometrically and practically).
The Present Study

In the present study, 321 first and second-level managers from another large Canadian utility company will be administered the same in-basket exercise used in the preliminary studies (1 and 2), but a newly-developed Insight questionnaire designed to measure the candidate's understanding of the situation is included in the exercise. Criterion data of supervisory ratings of performance will be collected concurrently. In order to cross-validate the findings from Study 2, the same scoring strategies and the identical item-subset compositions will be re-applied to this new sample, which will be scored for the same critical performance dimension evaluated in Study 2. The primary purpose of the present study, then, is to examine the degree to which the observed validities are spuriously inflated because of capitalization on chance, resulting in statistically significant correlations simply on the basis of chance. Assessment of the amount of shrinkage in the cross-validated coefficients will thus allow a more accurate determination of the true validity of the critical performance dimension selected for investigation, as measured by the scoring strategies developed in Study 2.

An additional goal of the present study follows from a finding of high intercorrelations among dimension scores in Study 2, as well as reported disagreement in the literature regarding the complexity of in-basket performance dimensions (Frederiksen, 1966; Lopez, 1966). For these reasons, a factor analysis of actions endorsed by candidates will be conducted in order to produce an empirically, rather than rationally, derived set of performance dimensions. A further aim of the present study is to assess the psychometric properties of the newly-designed Insight questionnaire. That is, to what degree will the aspects of in-basket performance measured by this new instrument improve the prediction of administrative ability?
A more detailed outline of the hypotheses and analyses involved in the present study will be provided in the fourth main section of this work entitled Rationale and Hypotheses for the Present Study. First, however, the conceptual and empirical groundwork for both the preliminary and present studies will be laid in the Literature Review. Next, the specific procedures and findings from the initial development and application of the in-basket used in the present study are related in the Preliminary Studies section. Then, overall findings from the Literature Review and Preliminary Studies sections will be integrated in the Rationale and Hypotheses for the Present Study section in order to provide a solid conceptual and empirical justification for the hypotheses and analyses in the present study. Finally, the procedures, findings, and implications of the present study will then be described in the Method, Results and Discussion sections.
LITERATURE REVIEW

The following Literature Review is divided into three sections. In the first section, management simulations and their relation to in-basket exercises will be explored in some depth. Section 2 then presents a summary of the published findings of the psychometric properties of the in-basket exercise (i.e., reliability and validity). The third and final section of the Literature Review provides an overview and empirical evaluation of the current approaches to scoring the in-basket, including those based on in-basket technology and scoring methods presently used in industry.

Section 1: Management Simulations

History and Development

One consequence of the varied influences on the history of management simulations has been considerable confusion regarding the definitions of both games and simulations (Biggs, 1990; Meier, Newell & Pazer, 1969). The lack of uniform terminology has resulted in little consensus regarding the definitions of and relationships among simulations, work sample tests, management simulations, management games, and in-basket exercises. The literature affords no clear classification system for these measurement instruments. Accordingly, a primary goal of this section is to provide a structure which clarifies the definitions of and relationships among these instruments. The terminology and relationships provided here may be more restrictive than those typically implied by researchers of games and simulations.
Terms and Definitions

The history and development of management simulations is closely tied to the history and development of games. A simulation is an imitative representation of real-world events in which the essential features of an activity are duplicated without portraying reality itself (Jones, 1972; Morgenthaler, 1961). In its broadest sense, a game is defined as an interactive process between the game system and one or more players who are given background information to study, rules and conditions to follow and are usually provided specific roles to play (Jones, 1972). Consequently, games that attempt to emulate real-world events or processes are also types of simulations, and the marriage of the two has borne a field known as gaming simulations (Jones, 1972).

In an organizational context, a managerial simulation attempts to mimic the essential features of managerial activity by recreating the typical work, decisions, and challenges of a managerial position. Management simulations are a broad classification of assessment devices which includes the management game as a smaller sub-type of assessment device. Management games simulate important aspects of business operations by involving individuals or groups of participants who partake in the running of important aspects of business operations and managerial functioning in a "live" fashion (Cascio, 1987; Howard, 1983). By reconstructing and representing critical business processes, the management game is therefore a type of managerial simulation and is properly classified as a type of gaming simulation.

The in-basket exercise has been classified as a management game (Lopez, 1966), and is thus also a gaming simulation. Lopez delineated three main classes of management games: solitaire games, small group games, and complex team games. The in-basket exercise is perhaps the most well-known example of the solitaire game, in which the player or candidate responds independently to a series of administrative or
general management problems as if he or she had assumed a particular manager's role. An example of a small group game is The Manufacturing Problem used in the 1956 Management Progress Study at the American Telephone and Telegraph Company (AT & T). In this game, a manufactured product was represented by tinker toys that could be built into objects of varying complexity (e.g., ladder, airplane). The challenges for the group of six participants were deciding what type of tinker toy product to manufacture, buying parts, and selling the finished product given changing market conditions (Bray, Campbell & Grant, 1974). An example of a complex team game is The Looking Glass (Lombardo, McCall & Devries, 1976) which involves 20 participants (placed in three divisions across four plant levels) who, in a day-long total organizational simulation, run a fictitious glass manufacturing company. Whether in the form of solitaire, small group or complex team games, management games continue to be widely-used for a variety of managerial functions, including assessment and diagnosis, training, and research into managerial behaviour (Faria, 1987).

While all management games are simulations, not all management simulations are management games. The leaderless group discussion, for example, is a simulation which involves a small group of participants who are asked to carry on a discussion, usually about a work-related topic (Bass, 1954). This simulation is characterized by a lack of structure and standardization; no person is designated as leader and there are no restrictions on the amount or content of participants' discussion. Individual interpersonal skills such as leadership, persuasiveness, flexibility, and verbal expression are the most commonly measured aspects of performance. Although roles may be assigned, the leaderless group discussion is not strictly a game in that it does not provide set rules and conditions for participants to follow, and it is not designed to simulate key aspects of business decisions and operations.
Management games, although properly classified as gaming simulations, are commonly referred to in the literature simply as simulations (Thornton & Cleveland, 1990). The considerable confusion in the definitions of games and simulations is exacerbated by the oversimplification and inconsistent use of these terms. In this work, the convention of using the shortened term, simulation, will be followed in referring to both gaming and non-gaming simulations. Accordingly, the in-basket exercise, although a management game, will be referred to more generally as a management simulation.

**Historical Background of Management Simulations**

Management simulations grew out of developments in four main areas: military war games, operations research, role playing, and performance testing (Thornton & Cleveland, 1990). War game simulations have a long history, beginning with their use in the re-enactment of battles in China in 3,000 B.C. for military education and training (Keys & Wolfe, 1990). In this century, war games have been adapted to reflect a variety of military situations ranging from actual physical battle simulations to more logistical, administrative aspects of military operations (Hausrath, 1971). An example of a simulation in military operations research is the 1955 Rand Corporation's Monopologs simulation of the U.S. Air Force supply system which provided decision-making experience without the risk of giving critical responsibilities to untrained staff (Jackson, 1959). The methodology of operations research was then applied to problems of management in non-military settings. For instance, Monolpologs led to the development in 1957 of the first widely-known management game, Top Management Decision Simulation, by the American Management Association (Ricciardi, Malcolm, Bellman, Clark, Kebbee, & Rawdon, 1957). The third development, the role-playing exercise (Moreno, 1959), allowed greater realism for those simulations that involve the adoption or enactment of certain roles. The fourth
and perhaps most influential development in the history of managerial simulations was the large-scale application of managerial simulations as a performance test, which is a standardized assessment of what candidates can do, rather than what they know, their knowledge being commonly measured by paper-and-pencil tests (Cascio & Phillips, 1979). The important step toward the use of management simulations as a performance test in personnel assessment deserves closer inspection, for it was here that the first business in-basket was developed and applied.

Historically, because of the central role of paper-and-pencil tests (rather than performance tests) in personnel assessment and selection, there was limited opportunity for the measurement of behaviours or actions. However, during the 1940's, German military psychologists began to use multiple assessment procedures which allowed the candidate to show behaviours in complex situations in order to generate a holistic appraisal of abilities, rather than an atomistic appraisal determined by the paper-and-pencil approach (Cascio, 1987). During the Second World War, the U.S. Office of Strategic Services (OSS) adopted the holistic approach in its use of multiple assessment techniques to select spies. A principal technique used by the OSS was the situational test, which is based on the belief that behavioural performance results from the interaction of both individual and situational variables (Cascio, 1987). Situational tests, then, are reconstructions of typical, realistic situations which are representative of the performance to be predicted (Flanagan, 1954). Complexity is a key feature of the situations presented so that candidates are less likely to discern the specific reactions or variables being measured. It is believed that allowing candidates to behave naturally and spontaneously will yield more typical, valid data than could be generated by other, less realistic measures. The OSS required potential spies to develop a cover story to hide their identity; several complex situational tests were then administered to try to trick candidates into breaking their cover (OSS, 1948).
As noted, the situational test, hallmark of the OSS assessment approach, is designed to re-construct or simulate real-world scenarios or events. Situational tests, then, can also be called simulations. The situational assessment principles of the OSS were adapted for managerial selection and were first applied in Douglas Bray's initiation of the 1956 Management Progress Study at AT & T. This longitudinal study is described as the largest and most comprehensive examination of managerial career development ever undertaken (Cascio, 1987). The Management Progress Study is also the first known industrial application of the assessment centre method which is characterized by multiple assessment procedures and pooled judgements of multiple assessors to evaluate the performance of managerial candidates. Using multiple assessment techniques, the primary purpose of the study was to identify those skills and characteristics that were most predictive of managerial potential. The assessment program (i.e., battery of measures used to assess candidates) included "standard" assessment procedures such as paper-and-pencil tests, interviews, and projective tests. However, the inclusion of situational tests in the assessment program was an innovative feature of the Management Progress Study. A variety of simulations were introduced, including the in-basket exercise.

As noted in the Introduction, the in-basket had been developed in the 1950's as a training tool for officers in the U.S. Air Force (Frederiksen et al., 1957), but was readily adaptable to a variety of settings and applications. John Hemphill, then at Educational Testing Service (ETS), worked with AT & T on the design and materials for the first assessment centre application of the in-basket. As a result, what is believed to be the first business in-basket exercise was developed by ETS, in conjunction with AT & T, for the assessment program of the Management Progress Study (Crooks, 1974). Norman Frederiksen, whose pioneering work introduced the in-basket as an effective training instrument, was also a research associate at ETS. Thus,
the early developmental and experimental work by Hemphill and Frederiksen established ETS as the founding figure in the history of the in-basket.

Management Simulations as Work Sample Tests

The Introduction noted that the in-basket can be considered as a member of a broader category of instruments known as work sample tests, which present samples of work representative of the target job and measure candidates' ability to do the work presented in the test. The in-basket is also a type of managerial simulation, which mimics the essential features of managerial activity by recreating the typical work of a managerial position and requiring the candidate to take action on the problems presented. By providing typical examples of managerial work, management simulations, then, are also a type of work sample test. Several figures will be provided in the following discussion to illustrate the relationships among these instruments. Because the literature more commonly refers to situational tests as simulations (it will be recalled that these terms are interchangeable), this convention will be followed here.

The area of work sample testing has undergone extensive growth in response to legal and social pressures regarding the fairness of selection procedures. Employment discrimination is minimized with the use of work sample tests because, as will be shown, they are well-suited to reducing the possibility of bias or adverse impact (Cascio, 1987). A test is biased if consistent non-zero errors of prediction are made for members of a subgroup (Cleary, 1968). Adverse impact is a condition in which a substantially different rate of selection in hiring, promotion, or other employment decisions work to the disadvantage of members of a race, sex, or ethnic group (Schneider & Schmitt, 1986). How work sample tests (and, by extension, managerial simulations and in-basket exercises) reduce the possibility of bias and adverse impact
will now be considered by looking more closely at the theoretical underpinnings of this
important class of instruments.

Historically, measurement for predictive purposes has been founded on the
notion that test results are "signs" or indicators of predispositions to behave in certain
ways. A different and perhaps more compelling view was provided by Wernimont and
Campbell (1968), who argued that prediction of behaviour would be most fruitful if
"samples", rather than signs, of behaviour were studied. Wernimont and Campbell's
views have been termed more generally as the "behaviour-consistency model" because
they argued that predictor measures (e.g., selection tests) should be as similar as
possible to outcome or criterion measures. As a result, in order to understand and
predict behaviour in organizations, tests which are related to observable job-behaviour
measures should be used. As will be discussed in more detail, work sample tests
provide an accurate "past performance indicator" by presenting actual or imitative
work samples. Therefore, they comport with the basic notion of the behaviour-
consistency model that the best predictor of future performance is past performance.

Asher and Sciarrino (1974) conducted a review of work sample tests, resulting
in the construction of a two-category classification system of work sample tests: a)
motor tests involving the direct manipulation of objects and b) verbal tests presenting
problem situations that are mainly language or people-oriented. Two examples of a
motor test are a typing test for office personnel (Giese, 1949) and a meat weighing test
for meat scalers (Bridgman, Spaethe & Dignan, 1958). Verbal work sample tests can
be classified as either group discussions/decision-making tests or individual situational
decision-making tests. Examples of verbal tests include the leaderless group discussion
(Bass, 1954), a role-playing test that simulates telephone contacts with customers (Gael
& Grant, 1972), and the in-basket exercise. Because the in-basket exercise is usually
completed independently, it is properly classified as an individual situational decision-making test.

An important distinction arising from Asher and Sciarrino's (1974) classification system is whether the work samples presented in the test are actual versus imitative representations of work. This distinction is useful in understanding the relationship between work sample tests and simulations, illustrated in Figure 1. It will be recalled that, in its broadest sense, a simulation is an imitative representation of real-world events in which the essential features of an activity are duplicated without portraying reality itself. In terms of Asher and Sciarrino's (1974) work sample classification system, it is likely that motor work sample tests are not simulations because no imitation of reality is presented; instead, actual work (i.e., reality itself) is presented. However, it is likely that verbal work sample tests are also simulations because a representation or image, rather than the actual work, is being presented and candidates are required to make decisions similar to those made in the job in question. In such cases, then, the simulation is also a type of work sample test, but a simulated one; it is an imitative representation of work (Howard, 1983).

It would appear that all managerial simulations can be classified as verbal work sample tests. Not all verbal work sample tests are managerial simulations, however. The broader category of verbal work sample tests includes non-managerial simulations, such as a law school admission's test involving cases, data interpretation, and reading comprehension (Breslow, 1957), to cite one example. As shown in Figure 1, not all simulations are work sample tests. Non-work sample simulations are typically used to generate a sequence of activities in a system and record statistics regarding system operation. Thus, the computer-based construction and analysis of mathematical models (e.g., economic analysis) are common forms of non-work sample simulations.
Figure 1. Relationship between simulations and work sample tests.
Furthering the behaviour-consistency model of Wernimont and Campbell (1968), Asher and Sciarrino (1974) hypothesized that the closer the "point-to-point" correspondence between the predictor and criterion, the greater the validity. They compared the validities of several classes of predictor tests and found that work sample tests generally yielded higher validity coefficients than ability, aptitude or personality tests. More specifically, when job proficiency was the criterion (measured by supervisory ratings), motor work sample tests were a close second to biographical information in terms of the number of high validity coefficients. Why biographical information was most predictive of job proficiency was puzzling to Asher and Sciarrino. Verbal work sample tests were "in the top-half of predictors," yielding higher validity coefficients than personality, mechanical aptitude, spatial relations and finger dexterity tests in predicting job proficiency. When the criterion was success in training (measured by grades achieved or a rating), verbal work sample tests had substantially more significant validity coefficients than motor work sample tests. Comparisons of the validities of different tests in predicting training success were not reported.

Asher and Sciarrino's (1974) finding of the strong predictive power of work sample tests was echoed by Hunter and Hunter's (1984) meta-analysis of several predictors of job performance, including the work sample test. Concerns regarding the adverse impact of cognitive ability tests lead to their review of the cumulative research on the use of alternative predictors of job performance. As Hunter and Hunter acknowledged, "the use of cognitive ability tests presents a serious problem for American society; there are differences in the mean ability scores of different racial and ethnic groups that are large enough to affect selection outcomes (p. 73)." In contrast, concerns regarding the adverse impact of work sample tests are considerably less than for achievement or aptitude tests because differences in test scores between
majority and minority groups are substantially smaller and usually nonsignificant (Robertson & Kandola, 1982; Schmidt, Greenthal, Hunter, Berner & Seaton, 1977). Hunter and Hunter reviewed the reported validities for six predictors used for promotion or certification where supervisor ratings were used as the criterion for current job performance. The resulting mean validities of the predictors were close in magnitude, ranging from .43 to .54. Work sample tests emerged (marginally) as the best predictor, with an ability composite predictor a very close second at .53. The goal of finding a predictor that was as valid as ability but that resulted in less adverse impact appeared to have been met.

Hunter and Hunter (1984) further conducted a utility analysis (Brogden, 1949; Schmidt, Hunter, McKenzie & Muldrow, 1979) of the predictors in order to quantify the dollar value gains in productivity with the use of various methods of selection. They determined that, when used for promotion or certification decisions, the gross utility estimates of ability and work sample tests were essentially equal, with a productivity gain for the U. S. federal government of $15.61 billion for one year due to hiring on the basis of ability compared to a productivity gain of $15.33 billion due to hiring on the basis of work sample tests (net utility estimates were not provided). However, they acknowledged that, unlike ability tests, work sample tests incurred high costs that were not taken into account in the utility calculations (e.g., initial development and re-development costs as the job changes, and high administration costs). Practical considerations, therefore, may limit the applications suggested by the favorable empirical findings reported for work sample tests.

In concluding this discussion of management simulations as work sample tests, it should be clearly understood that all management simulations are also work sample tests. We recall that the in-basket exercise is a type of management simulation. Thus, the in-basket is also a type of work sample test. As such, the in-basket exercise enjoys
the advantages of reduced bias and reduced adverse impact, but it also suffers the disadvantage of reduced utility (both gross and, in all likelihood, net) because of high development and high administration costs. Figure 2 provides an overview of the relationships among simulations, work sample tests (including the motor versus verbal distinction), management simulations, management games, and in-basket exercises. It should be clearly recognized that, in the literature, the similarities and differences among these instruments are not clear, and so the paradigm presented here is one of the unique contributions of this study.
Figure 2. Relationships among simulations, work sample tests, management simulations, management games, and in-basket exercises.
Issues in Management Simulation Design and Research

Simulations vary greatly in the nature and complexity of the stimuli (background information, items, etc.) and the nature and complexity of the response format. The traditional form involves written stimulus materials and open-ended written responses. A recent trend in management simulation design is toward the use of videotaped stimulus information and either open-ended written responses or multiple-choice responses (Goldsmith, 1991). Another new development in the use of simulations that are completely computer-based. Here, scenarios and situations are presented via microcomputers and the candidate responds using the computer keyboard. For example, in a complete computer-based simulation exercise called Utopia, the candidate is given the task of governing a fictitious island by simultaneously stimulating the economy and protecting the environment during his or her term (Diete, 1991). Currently, the majority of simulations used in assessment centres are non-computerized. For this reason, the following discussion of design and research issues concerning management simulations will be limited to non-computerized simulations.

Key Theoretical Issues: Realism and Fidelity

Realism. In-baskets are designed to provide a setting that closely approximates reality by requiring participants to "appreciate the social subtleties and technical niceties that always complicate any management problem" (Lopez, 1966, p.68). In this way, realism is a key feature, which is increased by providing ambiguous rather than straightforward issues and by presenting problems that are meaningful and appropriate given the background of participants and the target job (Gill, 1979; Lopez, 1966). Realism is also enhanced by having participants feel time-pressured, which requires them to use judgement in deciding which items and issues are most important, and
more closely simulates the action-oriented, quick-thinking approach needed for most managerial work (Lopez, 1966).

Realism is necessary to provide motivation for the participants to take the exercise seriously and become "ego-involved" (Lopez, 1966). When they can identify with the prescribed roles, participants become more involved and they are more likely to behave in the exercise as they would under similar circumstances in the organization (Wernimont & Campbell, 1968). In this way, the key situational test principle of allowing participants to behave naturally and spontaneously in order to yield more typical, valid data than could be generated by other, less realistic measures is supported.

It must be recognized that increasing realism to augment the richness of the information provided by the in-basket can present difficulties, particularly in administration and scoring. Either or both the stimulus materials and response materials can be targeted for greater realism. Improving realism of the stimulus materials by introducing "spontaneous" announcements of budgetary changes, lost shipments, etc. can reduce the control and standardization of the exercise, as administrators must ensure the announcements come at precisely the same point across administrations. In addition, making the response materials more realistic by providing several types and sizes of stationary usually makes both preparation of the exercise packages and scoring of the mass of papers more time-consuming and more costly for the organization.

**Fidelity.** A more precise, operational definition of realism applied to assessment exercises has been called the *fidelity* of the exercise, or the degree to which the task stimulus and response format match the conditions on the job (Thornton, 1992). Fidelity is viewed as a continuum of realism, with a decrease in fidelity as
stimulus materials and responses become less and less exact approximations of job stimuli and responses (Motowidlo, Dunnette & Carter, 1990). The most technical, widely-used definitions of low- and high-fidelity exercises were applied to simulations by Motowidlo et al. A high-fidelity simulation, they asserted, presents realistic, accurate samples of the task stimulus and elicits actual responses for performing the task. In contrast, a low-fidelity simulation would likely present a verbal or written description of a hypothetical work situation. In a low-fidelity simulation, participants describe or choose how they would deal with situations (typically in a questionnaire format) rather than directly handling problems.

Following Wernimont and Campbell’s (1968) logic of behaviour-consistency, Motowidlo et al. (1990) postulated that because high-fidelity simulations most closely resemble actual work conditions, high-fidelity simulations should be better indicators of future job performance than low-fidelity simulations. High-fidelity simulations, however, can be very costly to develop, implement, and score. Motowidlo et al. (1990) were unsure whether the gain in predictive potential from the use of high-fidelity simulations would justify their high costs. Accordingly, they developed a low-fidelity simulation for selection in order to explore the predictive usefulness of low fidelity.

A simulation was produced which presented 58 short, hypothetical problem situations (items) with a multiple-choice response format for each item. Five alternative courses of action were listed after each item; the participant was instructed to select the one alternative he or she would most likely take and the one alternative he or she would least likely take for each of the 58 task situations. The scoring key was developed from experienced managers' ratings of the most effective and least effective alternatives. Each of the two alternatives chosen by the participant was scored as either -1, 0 or +1, depending on its identification by the experienced managers as the
least effective, neutral, or most effective way to handle the situation. In a sample of 120 managerial incumbents, a correlation of .30 ($p < .01$) between total simulation scores and supervisory performance ratings of overall effectiveness was observed. Motowidlo et al. (1990) concluded that a carefully constructed low-fidelity simulation can yield satisfactory validity. In other words, the degree of realism provided in complex and costly high-fidelity simulations is not always necessary in order to obtain empirical validity. As a result, most organizations must carefully consider the choice between a low versus high-fidelity instrument, particularly in view of the current economic recession.

It should be noted that the term "fidelity" as it is used here is quite different from its original meaning, rooted in psychological testing and decision theory. Communications engineering provides the framework for describing the dilemma often faced by test developers, that of a choice between "wideband" and "narrowband" tests. The more varied and fast wideband signal transmits more information, but the clarity or dependability of the information received ("fidelity") is usually less than for the slower, homogeneous message sent along the narrowband. In general, narrowband signals have greater fidelity, with fewer errors confusing the signal. In this context, then, fidelity refers to the thoroughness of testing to obtain more certain information (Cronbach & Gleser, 1965), rather than the more recent adaptation to describe the realism and complexity of simulations.

The clarification of the terms bandwidth and fidelity (both in the broader, traditional test theory and the more specific, recent application to simulations) is important, for it will provide a structure for classifying and describing different simulations, including in-basket exercises. That is, a simulation designed to measure a broad range of behaviours may be seen as a wideband exercise. Conversely, a simulation which focuses on one key aspect of managerial behaviour can be seen as a
narrowband exercise. In addition to the bandwidth continuum, the level of fidelity can also provide a means of classifying simulations. That is, both wide and narrowband exercises can vary in the degree of realism in the task stimulus and the manner of response. As we shall see, high-fidelity, wideband simulations likely would pose greater challenges in establishing psychometric soundness and predictive usefulness, as well as presenting practical disadvantages in terms of lengthy scoring time.

Validity in the Context of Management Simulations

The reliability and validity of managerial simulations can be evaluated in order to determine their effectiveness as assessment devices. The psychometric principles underlying the reliability of managerial simulations are relatively straightforward and easily extend to in-basket exercise (to be described in Section 2 of the Literature Review). In contrast, the principles behind the validity of managerial simulations are more complex; clarification of these psychometric underpinnings will make the forthcoming extension of the psychometric principles underlying the validity of managerial simulations to in-basket exercises more comprehensible.

Validity, or the extent to which a measurement procedure does measure what it is designed to measure, can be assessed by looking at three main sources of evidence, namely, content-related evidence, construct-related evidence and criterion-related evidence (American Psychological Association Standards, 1985). Validation is the process of gathering and evaluating data to examine these sources of evidence. Two key goals of validation are to determine what the instrument measures (i.e., constructs) and how well it measures those constructs (Cascio, 1987). Because legal imperatives concern the fairness of selection measures, the following discussion of validation strategies for managerial simulations will be concerned with those instruments used for selection (rather than training, for example).
Content validity. In general, the content validity of measurement procedures refers to whether the procedures contain a representative sample of the universe of situations they are intended to reflect (content domain). Content validity is a key psychometric foundation of selection procedures; more specifically, it refers to the degree to which the content of a selection device is representative of important aspects of job-related performance (Uniform Guidelines on Employee Selection Procedures, 1978). Well-designed management simulations typically involve some type of job analysis (critical incidents, interviews, etc.) in order to determine the central tasks and requirements of the target job and incorporate them into the assessment device. Depending on their purpose, management simulations may be designed to measure either broad content domains (e.g., general administrative ability) or more narrow content domains (e.g., specific decision-making styles of administrators) through the use of either wideband or narrowband instruments. According to Thornton and Cleveland (1990), management simulations should compare very favorably in content validity with more traditional paper-and-pencil tests because of the careful attention typically given in the design and presentation of job-related stimulus materials and realistic response formats.

Construct validity. As applied to selection, construct validity refers to the extent to which a selection measure can assess candidates' levels of identifiable characteristics which have been determined to be important for successful job performance (Uniform Guidelines on Employee Selection Procedures, 1978). The relation between the targeted content domain and the features of a measurement instrument is an important consideration in establishing construct validity (Thornton & Cleveland, 1990). In other words, in order to ensure construct validity, consideration must be given both to the targeted managerial skills and to the type of assessment device chosen to measure those skills; some constructs cannot be adequately measured
by instruments other than a simulation. For example, some essential skills involved in
general managerial behaviour include interactions with personnel, demonstrating a
sound decision-making process, and showing leadership, to name a few. Simulations,
by allowing written role-play interactions with fictitious characters, may be more
appropriate than non-situational, traditional paper-and-pencil tests for the measurement
of those general managerial skills that are social or interactive in nature. As Thornton
and Cleveland (1990) pointed out, "simulations may be necessary to engage social
processes and to measure the application of social skills" (p. 195).

Perhaps an important qualifier should be added to this statement; high-fidelity
simulations may be necessary to engage social processes and measure the application of
social skills. High-fidelity simulations, with their open-ended response format, allow
for flexibility, spontaneity, and individuality which are necessary to adequately measure
socially-oriented skills. On the other hand, low-fidelity simulations, with their
relatively unrealistic stimulus items and restricted, multiple-choice response format,
may not be capable of tapping into the social process constructs which are important
aspects of general managerial skills. The multiple-choice, questionnaire format of low-
 fidelity simulations can measure knowledge and beliefs about social interactions, but
not the direct application of social skills through interactions with fictitious characters.
Low-fidelity simulations, therefore, may be better-suited to measure more narrowly-
defined, specific managerial behaviours that are not socially-oriented.

In sum, because general managerial skills, by their nature, include social skills
(determined to be necessary for successful job performance), an instrument which is
capable of measuring those interactive skills is necessary in order to achieve a high
degree of construct validity. Therefore, we may expect that construct validity would be
greatest when realistic, high-fidelity simulations are used in the measurement of general
managerial performance.
**Criterion-related validity.** Evidence of criterion-related validity is demonstrated by selection measures shown to be predictive of or significantly correlated with important elements of work behaviour (Uniform Guidelines on Employee Selection Procedures, 1978). Criterion-related validation is the most appropriate and most important validation procedure to apply when measures of individual differences are used to predict behaviour. Accordingly, the psychometric principles underlying the criterion-related validity of selection measures in general and managerial simulations in particular will be considered in some detail.

Two alternative approaches to assess criterion-related validity are available, namely, predictive or concurrent validation. If both the predictor scores and the criterion scores are available and considered at the same time, concurrent validity is being assessed. If the criterion results are not gathered until some time after predictor data are gathered, predictive validity is being measured. In essence, concurrent validation studies are concerned with assessing existing level(s) of behaviour(s) whereas predictive validation studies attempt to predict future performance level(s) of behaviour(s). Both approaches are primarily concerned with assessing the strength of the predictor-criterion relationship.

The most appropriate criterion-validation procedure (concurrent versus predictive) must take into account the purpose of measurement. If an instrument is to be used to make predictions of administrative ability, for instance, it would be most appropriate to conduct a predictive validation study using a longitudinal research design. In such a study, a typical sequence of events would be as follows: (a) assess administrative ability for job-candidates, (b) select candidates without using the administrative ability results, (c) gather criterion performance data some time in the future, and (d) measure the strength of the predictor-criterion relationship. However, real-world conditions of economic/budgetary constraints, high rates of staff turnover,
and so on, usually preclude the feasibility of predictive validation studies. Consequently, concurrent validation studies, where predictor scores are correlated with criterion measures for employees already on the job, are often substituted for predictive validation studies (Thornton & Byham, 1982).

Little published empirical information is available on most managerial simulations used in industry (Thornton & Cleveland, 1990) and, as a result, evidence of criterion-related validity for managerial simulations remains comparatively scarce. Considering the popularity and widespread use of managerial simulations, the paucity of published empirical validity may seem surprising. However, a brief consideration of the dynamic, heterogeneous nature of managerial simulations will help explain the unexpected lack of published findings.

First, it is clear that work sample testing (and, by extension, managerial simulations) is a dynamic field characterized by recent growth and responsiveness to rapidly changing job requirements. Unlike more traditional paper-and-pencil tests, managerial simulations are capable of incorporating new technologically-based design features, such as videotaped stimulus materials and computer response formats, for example. Moreover, in comparison to more static, homogeneous assessment measures (e.g., cognitive ability tests, personality inventories), management simulations are often more heterogeneous in both content (breadth of target behaviours: e.g., leadership ability, written and oral communication skills, personality characteristics) and form (e.g., leaderless group discussions, complex team games, etc.). This heterogeneity often results in company-specific, custom-tailored managerial simulations with comparatively limited, intra-company use. This limited use, in turn, translates into limited diffusion of costs for validation studies within industry, often making direct methods of establishing criterion validity (on a per exercise basis) cost prohibitive.
More indirect evidence of the criterion-related validity of managerial simulations comes from the studies of work sample validities noted earlier (Asher & Sciarrino, 1974; Hunter & Hunter, 1984; Roberston & Kandola, 1982; Wernimont & Campbell, 1968). It may be that, because many work samples are also managerial simulations, the evidence of work sample criterion-related validity has been generalized to apply to managerial simulations. However, as shown in Figure 1, not all work sample tests are managerial simulations. Moreover, given the widespread use of heterogeneous, company-specific management simulations, the generalization of criterion-related validities from broader work sample tests to particular, custom-tailored management simulations seems tenuous, at best.

Summary of Section 1: Management Simulations

The first section of the Literature Review outlined the development of the management simulation and acknowledged influences from several fields, including gaming. Because the literature affords no clear classification system, a primary purpose of this section was to provide a structure which clarified the definitions of and relationships among management games, management simulations, work sample tests and in-basket exercises. The in-basket exercise was classified as a type of gaming simulation. The historical background of management simulations was briefly related, followed by a discussion of management simulations as work sample tests. Conceptual issues and psychometric findings for work sample tests were reviewed. The in-basket exercise, as a type of work sample test, was seen to enjoy the same advantages (reduced bias and adverse impact) but also suffer from the disadvantages (reduced utility) of work sample tests. Issues in the design and research of managerial simulations were then explored. The notion of fidelity of simulations was introduced and discussed, followed by a brief consideration of realism in in-basket design and research. Lastly, the validity of managerial simulations was considered by looking at
the three validation strategies (content-, construct- and criterion-related validation) as they apply to management simulations.

Section 2: Psychometric Properties of the In-Basket Exercise

The Seminal In-Basket Work

The pioneering work by Frederiksen and his colleagues in the early 1950's established the founding principles and procedures that became the basis for conventional in-basket design and traditional objective scoring strategies. A synopsis of this seminal work, designed as a training device for officers in the U. S. Air Force, will illustrate several central issues of in-basket design (and subsequent development) and will set the stage for the review of the psychometric properties of the in-basket exercise which is to follow. This review of published information will be presented chronologically and will include a description of in-basket design and procedural developments, as well as a summary of empirical findings.

Goal of Frederiksen's work. Although this was not the first attempt to design a situational test measuring high-level performance, the in-basket developed by Frederiksen et al. (1957) was one of the first efforts to design a group-administered instrument that was presented in written form to measure individual performance. Frederiksen et al. recognized that, in the design of instruments to measure performance in high-level jobs, test developers usually face a choice between an objective, reliably scored instrument or one which is wideband or "sensitive" (i.e., can adequately measure the broad, complex set of skills involved in high-level jobs). A primary goal of Frederiksen et al. was to "devise a sensitive measure which may at the same time be objectively and reliably scored ....[which].... proceeds from the faith that progress
toward both goals of sensitivity and of objectivity may be made in one operation" (p. 2).

**Identification of performance dimensions.** The first step toward an objective and reliable scoring method was to identify those performance dimensions believed to be important aspects of effective executive performance. By studying the officers' training curriculum and by discussing with instructors the desired skills to be evaluated in the newly-trained Air Force officers, an initial set of 12 "functional categories of behaviour" (i.e., performance dimensions) was established. Four categories were selected from the 12 as the primary focus in the development of the in-basket, namely Flexibility, Efficient Use of Routines, Foresight, and Evaluating Data Effectively.

**Development and selection of exercise problems.** In the next step toward the development of an objective scoring system, the specific problems presented in the exercise were formulated in consultation with officers of several U. S. Air Force bases. Interviews were conducted and written descriptions of typical problems were prepared. In addition, the actual contents of administrators' in-baskets were examined. The final set of problems or items selected were designed to elicit one of the four performance dimensions listed above. The assignment of more than one problem to a particular performance dimension ensured that dimension was adequately measured. The problems chosen reflected the varying scope and complexity of administrative work, with simpler problems merely describing background information and other more complex problems (sometimes involving more than one item) describing more detailed administrative situations. A guiding principle in problem preparation was to minimize the amount of reading required by the participant, and so materials were designed to be brief and clear in an attempt to equalize reading comprehension skills required for optimal performance.
In order to minimize possible confounding effects of candidates' previous knowledge and/or experience with a similar situation, four two-hour in-basket exercises were constructed which required candidates to play four hypothetical roles in succession. In the first two-hour exercise, the required role was that of a Commanding Officer of a hypothetical Air Force wing. This was immediately followed by the role of the Director of Material in the second exercise, and the roles of Director of Personnel and Director of Operations in the third and fourth exercises, respectively. A comprehensive package of materials describing the hypothetical situation was provided to each candidate for each of the four exercises. Each package included detailed background information, an organizational chart, maps, a history of the particular division (e.g., Material, Personnel), and a mission statement for each division. In addition, stationery (letterhead, memo forms, paper clips, etc.) were furnished to encourage realistic responses. The broad scope of the package and the realism of the response materials were designed to stimulate candidates' complete involvement in the complex situations presented.

**Development of the scoring method.** The series of in-basket exercises were then administered to a "tryout group" of students training to be officers. Their responses to each problem were examined, and a list of the range of responses was derived by breaking down answers into the smallest, distinct units of action, later known as "action elements" (e.g., concurring in a recommendation, referring the problem to a higher authority). Only those action elements that were relevant to the predetermined functional category of behaviour (or performance dimension) for that problem were recorded. In other words, an action element may have primarily displayed Flexibility, but if the particular problem was designed to measure Foresight, the action element demonstrating Flexibility was ignored.
The entire list of action elements across all problems was then shown to two panels of expert judges in the Air Force who evaluated each action element and assigned scoring weights for the functional category of behaviour using a 5-point scale. Each point on the scale was designed to be used at least once for each problem. Using the judgements from the two panels, final scoring weights were then assigned by the test developers for each action element so that the scoring weights reflected a single functional category of behaviour. The panel of judges also ranked the importance of the problems presented in the exercise on a priority basis. The level of priority given each problem was later used in assigning scoring weights in the event a candidate failed to respond to the problem.

Significance. With some later variations, the process of striking a panel of expert judges to assign scoring weights for action elements became the established procedure in in-basket scoring key development for those researchers affiliated with Frederiksen and his colleagues at ETS. Although the exercise constructed by Frederiksen et al. (1957) was based on a situation involving the Air Force, the in-basket exercise was designed to be readily adapted for other situations and scenarios. Accordingly, as indicated in Section 1, the in-basket was soon adapted for its first industry application in the 1957 Management Progress Study. In the 1960's, the Port of New York Authority carried out extensive research and development with the in-basket exercise after adapting the problems and materials to reflect the nature of their work (Lopez, 1966). The Port Authority also adopted and further refined Frederiksen et al.'s expert panel process in the development and investigation of psychometrically-based objective scoring keys (this research will be explored more fully in the following review of empirical findings.)

The goal and approaches used in Frederiksen et al.'s (1957) seminal work have been considered here in some detail, not only for their historical significance, but also
because of their considerable influence on the present work. As noted earlier, the present work is a cross-validation of several scoring strategies, one of which is the expert panel approach initially developed by Frederiksen et al. In addition, the challenge originally identified by Frederiksen et al.--to design a situational instrument with the dual features of sensitivity and objectivity--has been a driving conceptual force behind the design and evaluation of the in-basket exercise investigated in the present work. However, Frederiksen's challenge has required some modification to its original two-part formulation. Specifically, a third aspect of practicality must be added to sensitivity and objectivity because the current global recession makes cost effectiveness a paramount concern. Practical issues of training time, scoring time and administration costs demand consideration in the overall evaluation of the in-basket exercise. The present work is an attempt to design and cross-validate an instrument which meets this three-part goal of sensitivity, objectivity, and practicality.

We turn now to a review of the reported empirical findings of the in-basket exercise. An examination of the published psychometric properties of the instrument will tell us whether Frederiksen's "...faith that progress toward both sensitivity and objectivity may be made in one operation..." (Frederiksen et al., 1957, p. 2) was justified.

Introductory Comments on In-Basket Empirical Findings

Over the last forty years, the in-basket exercise has been evaluated with respect to its psychometric properties of reliability and validity. To provide some structure for these findings, the empirical results have been summarized and presented in Table 1. It should be noted that the framework for the following review of results is a variation of an outline provided by Schippmann et al. (1990).
Summarizing empirical findings for the in-basket exercise is difficult because of several sources of variation across studies. Firstly, the research setting and therefore the content of the in-basket (which often reflects the research setting) varies widely across studies; universities, the public and private sectors, the military, and the educational administration system have all been selected as settings for in-basket research. Even for studies conducted within the same setting, there may be much variation in the level and specificity of the particular target job. For example, in management settings, the in-basket exercise may be designed for entry-level, first/second level, mid- or high-level managerial positions and may be tailored to reflect general or specific target jobs. Secondly, there is great variation in the type of criterion measures not only across disparate research settings but also within a single research setting. Within management settings alone, for example, salary level, job/salary progression, and supervisory ratings are commonly-used criterion measures, each based on different metrics.

A third source of variation across studies has to do with the in-basket exercise itself. The exercises examined in the following review vary greatly in several significant ways, including the complexity and scope of the performance dimensions to be measured by the exercise (wideband or narrowband), the realism of the stimulus materials and response format (high-fidelity or low-fidelity), the exercise construction methods, as well as the method of scoring.

In addition to these sources of variation across studies, there is a lack of important detail in published studies regarding these sources of variation (Frederiksen et al., 1972; Schippmann et al., 1990). For instance, information regarding the development and description of outcome criteria is often not provided, and important facts concerning the scoring method used to evaluate the in-basket exercise are simply not offered.
### Table 1

**Review of Major Empirical Findings on the In-Basket**

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Range</th>
<th>Method</th>
<th>Range</th>
<th>Reliability/Validity</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frederiksen, Saunders &amp; Wand (1957)</td>
<td>IR</td>
<td>.62-.93</td>
<td>C</td>
<td>.02-.14</td>
<td>Grades</td>
<td>Exam scores</td>
</tr>
<tr>
<td></td>
<td>IR</td>
<td>.47-.94</td>
<td>C</td>
<td>.01-.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frederiksen (1962)</td>
<td>SH^a</td>
<td>-.92</td>
<td>C</td>
<td>-.05-.24</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALT</td>
<td>0-.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemphill, Griffiths &amp; Frederiksen (1962)</td>
<td>SH</td>
<td>.52-.97</td>
<td>C</td>
<td>-.05-.24</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td>Lopez (1966)</td>
<td>IR</td>
<td>-.20-.97</td>
<td>C</td>
<td>-.12-.20</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td>-.15-.18</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td>-.29-.27</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td>Bray &amp; Grant (1966)</td>
<td>IR</td>
<td>.92</td>
<td>P</td>
<td>-.19-.44</td>
<td>Salary progress</td>
<td>OAR</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td>.45-.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross (1969)</td>
<td>SH</td>
<td>-.39-.95</td>
<td>P</td>
<td>-.45-.59</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALT</td>
<td>-.45-.75</td>
<td>C</td>
<td>-.33-.72</td>
<td>Supervisor ratings</td>
<td></td>
</tr>
<tr>
<td>Wollowick &amp; McNamara (1969)</td>
<td>P</td>
<td>.32</td>
<td></td>
<td></td>
<td>Job progression</td>
<td></td>
</tr>
</tbody>
</table>

Continued next page
Table 1 (cont.)

Review of Major Empirical Findings on the In-Basket

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Range</th>
<th>Method</th>
<th>Range</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meyer (1970)</td>
<td>SH</td>
<td>.50-.95</td>
<td>C</td>
<td>-.19-.37</td>
<td>&quot;Supervision&quot; rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>-.09-.31</td>
<td>&quot;Planning&quot; rating</td>
</tr>
<tr>
<td>Bourgeois &amp; Slivinski (1974)</td>
<td>IR</td>
<td>.54-.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brass &amp; Oldham (1976)</td>
<td>IR</td>
<td>.64-.95</td>
<td>C</td>
<td>-.17-.40</td>
<td>Supervisor ratings</td>
</tr>
<tr>
<td></td>
<td>SH^a</td>
<td>.13-.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hinrichs &amp; Haanpera (1976)</td>
<td>ALPHA</td>
<td>-.04-.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kesselman, Lopez &amp; Lopez (1982)</td>
<td>SH</td>
<td>.83</td>
<td>C</td>
<td>.13-.33</td>
<td>Supervisor ratings</td>
</tr>
<tr>
<td>Turnage &amp; Muchinsky (1984)</td>
<td>P</td>
<td>.08</td>
<td></td>
<td></td>
<td>Supervisor ratings</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>.25</td>
<td></td>
<td></td>
<td>Career potential</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>.03</td>
<td></td>
<td></td>
<td>Promotions</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>.01</td>
<td></td>
<td></td>
<td>Salary progression</td>
</tr>
</tbody>
</table>

Continued next page
Table 1 (cont.)

Review of Major Empirical Findings on the In-Basket

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Range</th>
<th>Method</th>
<th>Range</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hakstian, Woolsey &amp; Schroeder (1986)</td>
<td>SH(^a)</td>
<td>.80-.84</td>
<td>C</td>
<td>.00-.33</td>
<td>Supervisor ratings</td>
</tr>
<tr>
<td></td>
<td>IR</td>
<td>.80-.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brannick, Michaels &amp; Baker (1989)</td>
<td>IR</td>
<td>.71-.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SH</td>
<td>.19-.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALT</td>
<td>.21-.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALPHA</td>
<td>.35-.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tett &amp; Jackson (1990)</td>
<td>IR</td>
<td>.65-.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALPHA</td>
<td>.41-.70</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Table 1 is adapted from Schippmann et al. (1990). IR = inter-rater agreement; SH = split-half reliability; ALT = alternate form reliability; ALPHA = Cronbach's alpha; C = concurrent validity; P = predictive validity; OAR = Overall assessment centre rating.

\(^a\)Corrected using Spearman-Brown formula.
Reliability of the In-Basket Exercise

Typically, the most direct measure of the reliability of an instrument is obtained by the test-retest method (i.e., administering the same instrument to the same group of subjects at two different times). Correlations of the two scores arising from this method allow computation of the "coefficient of stability" of the instrument. However, random error is often introduced with the use of stability coefficients because of variables which influence the performance of participants on one test administration but not the other (e.g., differences in testing conditions, mood of participant, etc.). Furthermore, in some situations, multiple test administrations are not feasible; this is certainly true for the in-basket exercise, where, because of logistical difficulties and high costs associated with multiple test administrations, assessment of test-retest reliability is generally not viable.

As a result, other methods have been used to assess in-basket reliabilities, including analyses of alternate forms of the same instrument, internal consistency estimates (i.e., split-half analyses and, to a lesser extent, coefficient alpha), and the degree of inter-rater agreement. Each of these methods will now be briefly introduced, followed by a review of the specific empirical findings for each reliability estimation method as applied to the in-basket. A more complete description of the design and methodology used in the more central studies will be provided in the following Validity section.

Alternate Form Reliability Estimation

Conceptual foundation. Alternate form reliability estimation is based on the premise that an instrument contains only a sample of selected items from a content domain. In theory, it is possible to select a different, interchangeable sample of items
in order to construct two or more different forms of the same instrument which contain
the same number of items at the same level of difficulty with non-significant
differences in means and variances across the forms. The correlation between scores
obtained from the two forms yields a reliability estimate called the "coefficient of
equivalence," which takes into account error due to different samples of items.
Generally, alternate form reliability estimation requires as many administrations as
there are numbers of forms. The administrations are typically conducted as closely
together in time as possible in order to minimize random error arising from the
probable increase in confounding factors over time.

**Empirical findings.** Hemphill, Griffiths and Frederiksen (1962) developed and
applied several forms of The Whitman School In-Basket exercise, in order to assess
ability in school administration. The alternate form reliabilities were reported by
Frederiksen (1962), who had also developed and applied the Bureau of Business In-
Basket. and compared these results with those from Hemphill et al.'s (1962) study of
school administrators. The reported median r's ranged from .25 to .38 using the four
alternate forms across 68 performance categories. Frederiksen (1972) again compared
several alternate versions of an in-basket exercise and found reliability coefficients of
.15, .17 and .27 for the three comparisons conducted. Brannick et al. (1989) reported
reliability correlations ranging from .21 to .43 across the five performance dimensions
measured by two alternate forms of an in-basket exercise.

**Summary.** Generally, reliability coefficients based on alternate forms of an in-
basket are not strong. One possible reason for the low coefficients is that the situations
presented in the various forms of the in-basket can be quite different. For instance, the
study by Frederiksen (1962) involved alternate forms of an in-basket exercise
describing both a school administration situation and a business situation. Reliability
coefficients from alternate forms based on the school situation were higher than those
coefficients based on alternate forms across both the school and business situations. Therefore, as expected, there was greater consistency in performance across those forms dealing with the same, as opposed to different, situations.

A major difficulty in interpreting alternate form reliabilities is that in-basket performance across forms has typically been evaluated by multiple raters, causing performance reliability to be confounded with rater (or scorer) reliability. Little published information is available regarding the proportion of variance contributed by in-basket scorers. Overall, the very limited data available on alternate form reliability estimation allows only a tentative conclusion of a lack of strong consistency across individual in-basket performances.

Alternate forms of instruments are recognized as costly and difficult to construct (Cascio, 1987). Moreover, in some situations, because of logistical difficulties, only a single administration of a test is possible, making neither test-retest nor alternate form reliability estimation the method of choice for in-basket reliability estimation. Consequently, other methods are used which are based on measuring the effect of different samples of items on reliability, namely, methods of internal consistency.

**Internal Consistency: Split-Half and Alpha Coefficients**

**Conceptual foundation.** As noted, in cases where only a single test administration is possible, reliance is placed on measuring the internal consistency of the instrument, which is most commonly assessed by the split-half and coefficient alpha methods. Split-half analyses involve several methods of separating the test into two equivalent halves and then correcting for double length by the Spearman-Brown formula. As with alternate form reliability estimation, the correlation between scores based on the two halves of the instrument (alternate forms) are interpreted as a coefficient of equivalence. In split-half analyses, a common method of creating two
equivalent forms is to split the test in half by computing two separate scores for each individual based on responses to odd-numbered items and responses to even-numbered items. Other methods of splitting the test include randomly selecting the items, or selecting items consecutively to form the first half with the remaining items making up the second half (e.g., items 1-10 in the first half and items 11-20 in the second half of a 20-item test).

Coefficient alpha (Cronbach, 1951) is an additional technique used to assess internal consistency that is based on an analysis of item variances and is considered to be a measure of homogeneity (rather than equivalence) as it indicates the degree to which items within a test are intercorrelated. The coefficient alpha estimates the average correlation of all possible half-splits of items within a test (Cascio, 1987).

Empirical findings. In the following review of split-half reliability findings, the particular split-half method used (i.e., odd-even, random) has been specified where it is known. Because each item in the in-basket used by Frederiksen et al. (1957) was designed to measure only one of the four performance dimensions, and because the total number of items measuring each category ranged from merely four to a maximum of seven items, split-half estimates of reliability were not directly calculated. Instead, a series of intercorrelations of the groups of items measuring each of the four performance categories were determined; these ranged from -.31 to .42. Later studies brought developments in item preparation and in-basket design (detailed in the Validity section) which, among other results, made the calculation of split-half reliability more direct.

Hemphill et al. (1962) reported odd-even split-half reliability coefficients for 68 stylistic categories of in-basket performance ranging from zero to .97. On the basis of these results, forty categories were retained for subsequent analyses; the reliabilities of
these categories ranged from .52 to .97, with a median $r$ of .78. It should be noted, however, that Hemphill et al. employed different raters for the different halves of the test and so consistency of performance across the two halves of the exercise was confounded with the degree of inter-scorer agreement resulting from the use of different raters. Consequently, these reliability results should be interpreted cautiously.

Frederiksen's (1962) analyses with The Whitman School In-Basket and the Bureau of Business In-Basket resulted in Spearman-Brown corrected split-half median correlations (across performance categories) which ranged from .47 to .56. A study by Cross (1969) also examined the Whitman School In-Basket, and, in two administrations of the exercise, odd-even split-half reliability coefficients of .57 and .49 were reported.

Meyer (1970) conducted an in-basket analysis at General Electric Company by constructing approximately 50 performance categories. Results showed that out of these 50 initial categories, only 27 demonstrated sufficient reliability for further analysis, with odd-even split-half coefficients of the reduced set ranging from .50 to .95. Like Hemphill et al. (1962), Meyer used different raters for the different halves of the test, and, for reasons provided above, these results should be interpreted with caution (Schippmann et al., 1990). In 1976, Brass and Oldham found odd-even split-half correlations ranging from .13 to .58 across six scoring categories designed to measure leadership. Stronger reliability results were seen by Kesselman et al. (1982), who reported a split-half odd-even reliability coefficient of .83 for a composite in-basket score. Hakstian et al. (1986) obtained corrected split-half reliability coefficients from .80 to .84 for the Content dimension of a two dimension exercise. Brannick et al. (1989) reported odd-even split-half reliability coefficients for five dimensions of the two alternate in-basket forms ranging from .19 to .61 (Form A) and .20 to .62 (Form B), with an overall median $r$ of .34.
Using the coefficient alpha to estimate internal consistency, Hinrichs and Haanpera (1976) reported an average alpha coefficient of .49 measured across 14 performance categories of an in-basket exercise. Brannick et al. (1989) computed alpha coefficients for six dimensions across two alternate forms of an in-basket exercise. The results ranged from .35 to .72, averaging to .51. Alpha coefficients were also calculated by Tett and Jackson (1990) for the six aspects of participative decision-making measured in their in-basket exercise. The alpha coefficients ranged in magnitude from .41 to .70, averaging .52 across the six performance dimensions.

Summary. Reported reliability estimates derived from the split-half method of assessing internal consistency vary greatly. The results reported by Kesselman et al. (1982) and Hakstian et al. (1986), when averaged, suggest that in-basket exercises can attain satisfactory levels of equivalence-based estimates of reliability (approximately .82). However, the median split-half reliability coefficient of .34 reported by Brannick et al. (1989) is significantly lower, in sharp contrast to earlier, stronger findings of Kesselman et al. and Hakstian et al. Given such a wide range in reported findings, it is difficult to conclude, with confidence, whether or not the split-half reliabilities of in-basket exercises are satisfactory. In contrast, less variability is seen in the alpha coefficient results reported. Despite greater convergence, the mean coefficient alpha across studies is .51; this modest degree of homogeneity does not clearly indicate that the selected samples of items result in solid reliability.

In an internally consistent instrument (indicated by high coefficients of equivalence), items are considered to be mutually equivalent. In general, evidence of inconsistency (indicated by low coefficients of equivalence) is interpreted as error variance arising from inconsistent sampling of the content domain (Cascio, 1987). Overall, both the split-half and alpha coefficients reported here may seem somewhat
baskets, the findings summarized here may appear less negative, with such features as the breadth and complexity of items, cases of low within-item variance (due, in large part, to low endorsement rates of courses of action), and the numerous unquantifiable influences that affect the participant's perception of the items, constituting some possible reasons for the modest internal consistency results observed.

Inter-Rater Reliability Estimation

**Conceptual foundation.** The assessment of inter-rater agreement measures the likelihood that two or more scorers will arrive at similar scores or ratings for an individual's performance. Calculation of inter-rater reliability allows an estimation of the errors of measurement attributable to scorer variance (Cascio, 1987).

**Empirical findings.** In the context of the in-basket exercise, inter-rater reliability was first examined by Frederiksen et al. (1957), who compared individual item scores and the total score assigned by three raters and expressed reliability in terms of the product-moment correlation coefficient between various pairs of raters (e.g., Scorer A vs. B; B vs. C, etc.). Reliability coefficients from .47 to .94 were seen. Lopez (1966) conducted a study with police lieutenants in the Port Authority of New York and reported inter-rater reliabilities for 78 performance categories ranging from -.20 to .97 with a median $r$ of .60. When the 78 categories were collapsed to eight more global performance categories, the correlational range measuring inter-rater reliability improved, ranging from .35 to .97 with a median $r$ of .80. Bourgeois and Slivinski (1974) reported a median $r$ of .86 across nine performance categories.

Two years later, Brass and Oldham (1976) reported inter-rater correlations ranging from .64 to .95 across the six dimensions measured. Brostoff and Meyer (1984) computed correlations for two raters in terms of four performance dimensions; the resulting correlations ranged from .84 to .91. Inter-rater reliability estimates for
two dimensions measured in a study by Hakstian et al. (1986) were .80 and .99. Brannick et al. (1990) also reported high inter-rater agreement within each alternate form of the in-basket measured, with coefficients ranging from .71 to .94 across the two forms. Lastly, Tett and Jackson (1990) published inter-rater correlations from six categories of in-basket performance ranging from .65 to .95 with an average reliability coefficient of .82.

**Summary.** The inter-rater reliability findings suggest that performance dimensions measured by the in-basket exercise can, for the most part, be assessed reliably by scorers. Across studies and exercises, the reported reliability coefficients converge to an average between .80 to .85. However, within in-basket exercises, there can be a considerable range in inter-rater values (e.g., .35 to .97 reported by Lopez, 1966). This intra-exercise variation may be the result of differing levels of specificity of the performance dimensions and the degree to which they can be clearly and objectively operationalized. For example, we would expect a dimension concerned with productivity (measured, perhaps, by counting the number of words written and/or number of actions taken) to be more reliably assessed than a dimension concerned with sensitivity in written responses (measured, perhaps, by the use of a courteous tone). The range in inter-rater reliability coefficients across studies may be due, in part, to the differences in scoring methods used (i.e., subjective, objective, combination), or the efficacy of scorer training (e.g., whether scoring manuals were available, the length of the training process).

**General Summary of In-Basket Reliability Results**

In summary, it appears that evidence of reliability of the in-basket has been modestly established. Alternate form reliability results are relatively low. Low coefficients of equivalence are likely due, in large part, to the comparison of in-basket
exercises which are based on different scenarios and situations. Because of practical
difficulties and costs associated with multiple administrations, reliability estimates
based on equivalence and homogeneity of items (i.e., internal consistency) rather than
stability (i.e., test-retest) are favored forms of reliability estimation. However, there
are great variations in the magnitude of the split-half reliability estimates reported for
the in-basket, and internal consistency reliability estimates are generally not strong
across studies. More promising results suggesting that the in-basket exercise can be
reliably scored are seen in the reported findings of the degree of inter-rater agreement.
Coding schemes that are designed to be more objective, with behavioural and coding
rule descriptions set out in coding manuals probably account for the considerable levels
of inter-rater reliability seen across studies. However, no coding scheme for a free-
format response in-basket can completely eliminate the inherent subjectivity and
judgement involved in scoring an in-basket exercise. All in all, considering the
evidence of the reliability of the in-basket, it appears that Frederiksen's (1957) original
goal of designing an instrument with the property of objectivity, reflected by in-basket
reliability results, has been adequately met.

Validity of the In-Basket Exercise

For several reasons, the following review of the validity findings of the in-
basket exercise is not completely exhaustive. Not only is the literature voluminous, but
the research reported in this area is also fragmented and fraught with methodological
and conceptual shortcomings (Schippmann et al., 1990). Accordingly, only those
studies which sufficiently report necessary descriptive (i.e., development and design)
and evaluative (i.e., validity) information are included in this review. An additional
consideration guiding the selection of studies discussed here is the significance of the
contributions made by the work. Generally, those studies which reported unique and
important developments in in-basket design or which involved large-scale applications
of existing methodology were selected over those studies which simply replicated previous approaches or which involved small-scale applications, precluding the generalizability of findings. For example, the innovative in-basket design and scoring methods behind the development of Frederiksen's (1962) Bureau of Business In-Basket were also used in the development of The Whitman School In-Basket in a much more elaborate study by Hemphill et. al (1962). Consequently, only the in-basket principles and their application in the latter study are discussed in this review.

Criterion-related Validity

Air Force in-baskets. In order to validate the in-basket exercises used in the seminal work by Frederiksen et al. (1957), they were administered to a class of 92 students at the U. S. Air Force Command and Staff School. The efficacy of the in-basket as a measure to evaluate military instruction was assessed using a total in-basket score (summed across the four previously-mentioned roles) as the predictor measure. Both scores on course grades and an external educational test were used as criterion measures. Validity coefficients of .25 with the external test and .15 with course grades were reported.

Frederiksen et al. (1957) concluded their investigation by acknowledging that the in-basket exercises they developed were disappointing in terms of their psychometric properties. Consequently, they recognized that the in-basket exercise, as it was then, was not useful for the selection of individual candidates. Instead, the exercise was useful solely for training and educational purposes. Although discouraging to the researchers, the results of the Air Force in-basket study laid an important foundation for continued research and development with the in-basket.

School administration in-baskets. Further work by Hemphill, Griffiths and Frederiksen (1962) resulted in the first evidence that in-basket exercise performance is
positively associated with on-the-job administrative performance. Using the newly-developed series of three Whitman School In-Baskets (Frederiksen, 1962), they undertook a study of the administrative performance of 232 elementary school principals (137 men, 95 women). The Whitman School In-Baskets differed from the Air Force in-baskets in both the preparation of in-basket items and the method of scoring. As was shown earlier, each item in the Air Force in-baskets was prepared to elicit only one preselected behaviour judged to be relevant to the officer training program. In contrast, each item in the school administration in-baskets was written to reflect a broad range of target behaviours that were representative of the work required in school administration.

Two general components of school administration in-basket performance were targeted for evaluation, namely, style and content. Content referred to what specific actions were taken (e.g., praising a teacher), whereas style referred to the manner in which, or how, actions were taken in handling an item (e.g., using a courteous tone). These two aspects of administrative performance were not seen as independent, for it would be unlikely that praising a teacher (content) would be done discourteously (style). The inclusion of the content scoring category is notable, for it is a particularly unique contribution of this study. Identifying what specific actions were taken allowed the important development, introduced in subsequent studies, of evaluating the appropriateness of actions.

In scoring for style, a total of 68 scoring categories were selected based both on observations of school principals at work and on theories of leadership. For example, a leadership theory of Hemphill (1958) involved concepts of initiation of structure and consideration which resulted in the in-basket stylistic categories of Initiates a New Structure and Courtesy to Subordinates. Other observation-based categories included Postpones Decisions, Asks Subordinate For Information, and Sets a Deadline. Each in-
basket item then was measured using each of the 68 stylistic scoring categories and the overall score for a certain stylistic category was the number of times a particular style was evidenced across all items from all four in-basket exercises (132 items in total).

Following odd-even split-half reliability analyses, 28 of the original set of 68 stylistic categories were eliminated, leaving a final set of 40 categories. A factor analysis of these 40 stylistic categories revealed eight stylistic factors: Exchanging Information, Discussing before Acting, Complying with Suggestions made by Others, Analyzing the Situation, Maintaining Organizational Relationships, Organizing Work, Responding to Outsiders, and Directing the Work of Others.

In scoring for content, rather than style, the major courses of action taken by the school principals were identified for each in-basket item (e.g., refers to secretary, communicates with superintendent's office). The number of courses of action derived for any one item was arbitrarily limited to ten (no reason was provided). A value of 0 or 1 was recorded for each course of action, with 1 indicating the selection or endorsement of a particular course of action for that item. Several methods were then considered in scoring the courses of action, or content component, of in-basket performance. An interesting, but unsuccessful, effort was made to derive content scores by evaluating the appropriateness the courses of action taken as solutions to the problem presented in the item. This approach was abandoned because of several unsuccessful attempts to obtain a consensus among qualified judges of the appropriateness of the courses of action for an item. However, this panel-based logical approach of evaluating the appropriateness of actions was successfully adopted in subsequent studies and became the leading method of scoring the content of in-basket performance.
Ultimately, two methods of scoring for content of responses were adopted in the school administration study using the categories of Imaginativeness and Organization Change. A score of 1 for Imaginativeness was applied to those courses of action identified by scorers to be good, creative ideas that go beyond the courses of action immediately apparent in the exercise. A score of 1 for Organization Change was applied to those actions which reflected the introduction of, or consideration of, changes in the policies, practices or procedures of the organization.

In addition to scoring for style and content, Hemphill et al. (1962) devised a third scoring component that was seen as more global and subjective than the other two, more analytical, components. This third component required scorers to make impressionistic ratings of school principals' performance using 21 pairs of adjectives (e.g., friendly versus aloof, logical versus intuitive, witty versus humourless) and to select the one adjective from each pair which best described performance in the in-basket. The score for each adjective-pair was based on the number of times the first adjective in the pair was checked by the scorer (further rationale was not provided). On the basis of reliability estimates, 10 of the 21 pairs of adjectives were retained for subsequent analyses.

An additional development in the school administration study was the first large-scale introduction of a "Reasons for Action" form, which required candidates, upon completion of the in-basket exercise, to state very briefly what was done for each item and why. The primary purpose for the inclusion of the form was to clarify the actions taken and the intent behind the responses made in order to facilitate more accurate interpretation and scoring of responses. For instance, 1 of the 68 in-basket style categories was Number of Items not Attempted. It was recognized that, just because a candidate did not actually write in response to the item, did not mean that no action was attempted. On the Reasons for Action form, the candidate may have clarified that
action was postponed because it was not an urgent problem. Postponing a decision, in this study, was considered a scorable response.

Several criterion measures were employed in the school administration study, including ratings by subordinates (teachers) and ratings by superiors (superintendents). Because of their critical importance in actual administrative situations, only the criterion-related validity results from performance ratings by the superiors of the principals will be related here. The principals were evaluated on 13 performance appraisal dimensions, such as Interest in Work, Ability to Get Along with Parents, and an Overall General Impression rating. Each performance dimension used a 5-point subjective scale, multiplied by a number reflecting the judge's level of confidence about the rating.

Correlations between the superiors' Overall General Impression rating and scores from 32 stylistic categories were calculated (no reasons were provided for eliminating 8 of the final set of 40 stylistic categories). Only 4 of the 32 correlations were significant at the .01 level. Composite scores for the eight stylistic in-basket factors (listed earlier) were also approximated and correlated with the Overall General Impression rating. Here, correlations ranged from -.05 to .24, with two factors—Discussing before Acting and Exchanging Information—reaching significance. Lastly, the impressionistic ratings of the ten adjective pairs yielded validity coefficients ranging from -.15 to .17 when correlated with the superiors' Overall General Impression (three pairs reached significance).

Overall, the criterion-related validity coefficients for the school administration in-basket exercises were low. However, important methodological developments in item preparation and scoring approaches (i.e., the notion of a content category,
Reasons for Action form) set the groundwork for a more psychometrically-based design and scoring approach researched in the Port of New York Authority studies.

The Port of New York Authority studies. An important series of studies by the New York Port Authority represented the first use of the in-basket exercise as a means of assessing candidates for promotion to higher levels of management. Unlike previous training or research applications, in-basket performance for candidates in the Port Authority studies held serious, personal consequences because of the application of performance results to selection and promotion. These studies have been summarized by Lopez (1966), who provided much descriptive, but little quantitative, information (Frederiksen et al., 1972).

Historically, the Port Authority was committed to using objective assessment methods for selection and promotion and was actively engaged in researching a number of assessment devices. In the late 1950's, they determined that their assessment program was not measuring important skills such as critical thinking, organization of ideas, and communication skills. In 1960, the Port Authority then contracted with ETS (specifically, Frederiksen and Hemphill) to develop an in-basket exercise which would measure these and other skills for the evaluation of police lieutenants for promotion. The conclusion drawn from this evaluation was that, although seen as an improvement to its assessment program, the in-basket exercise was too complex and too costly to score (fully trained scorers required three hours to evaluate each exercise). Accordingly, until a larger-scale application was required, the in-basket exercise would not be used.

Four years later, the Port Authority initiated a comprehensive management selection and promotion program for both junior- and middle-level management. This was seen as an excellent opportunity to conduct further research and development with
the in-basket. Consequently, in 1964, Lois Crooks of ETS developed two in-basket exercises that were designed to assess predicted success in administrative and facility management jobs, respectively. The completed in-basket exercises were scored by ETS, who employed 42 of the 68 stylistic performance categories generated in the school administration study (Hemphill et al., 1962). Eight stylistic scoring factors similar to those identified by Hemphill were derived from the 42 categories. In addition, a "content" category scoring key was successfully developed using a panel of subject matter experts to logically assign a weight of -1, 0 or +1 for appropriateness of each course of action (no further development details were provided). A Productivity category was also scored, based on the number of items attempted, the number of courses of action taken, and the number of words written. Finally, an impressionistic rating by the scorer of how well the candidate would perform on the job was also made.

The criterion measure used in the study was a composite score derived from the ratings of 10 supervisors, calculated for each candidate. In the sample of 58 administrative service candidates, no significant correlations between any of the in-basket categories (or factors) and supervisory ratings were seen. However, in the sample of 97 facility management candidates, the impressionistic rating, Productivity, Content and several stylistic factors were significantly related (at both the .05 and .01 levels) to the supervisory rating. Significant correlations ranged from -.29 to .27.

The following year, the Port Authority's management selection and promotion program was again administered. However, in 1965, a different in-basket exercise was used, rather than the ones developed for the 1964 program. The "Ama Company, Inc." in-basket had been developed several years before for the American Management Association. Lopez (1966), who was involved in its development, confirmed that this in-basket was based on standard ETS principles and had already been administered to
over 2,500 executives. In a departure from the traditional ETS scoring approach, an innovative feature of the Ama in-basket was the inclusion of a multiple-choice, self-report questionnaire used to score the exercise. The in-basket "Action Report Form" was developed as a way to efficiently and economically meet the significant scoring demands of large-scale in-basket applications. Upon completion of the exercise, the candidate was asked to respond "yes" or "no" to a total of 822 statements provided to describe all possible actions which could be taken for all problems. Little provision was made for "unusual" responses because it was believed that, with statements based on tabulated responses from 2,500 participants, it would be unlikely that subsequent participants would respond in a significantly different way. A sheet was provided for unusual responses to be recorded, if necessary, but it appears they were not scored.

Few details are available regarding the scoring system developed and applied in scoring the Action Report. It is known that initially four in-basket performance factors (Judgement, Output, Social Style and Leadership), based on 12 stylistic categories, were scored. Analyses of the intercorrelations among categories and factors from a combined sample of Port Authority candidates and management executives from the 1965 American Management Associations' Management Course (totalling 726 participants) led to refinements of the performance categories. Specifically, a final set of three factors called Organization of Work, Productivity, and Delegation were identified and retained. It is not clear, however, how the factors were derived or how the 822 yes/no statements were weighed or assigned to the categories or factors. Lopez (1966) simply stated that the Action Report used a scoring system that was based on ETS procedures and previous research.

A complication with the reporting of this research was that, in the validation of the Action Report, the in-basket performance results and criterion data from 150 Port Authority candidates were combined with results from a study of 93 management
executives. The criterion measure used was a composite score, summed across seven job performance dimensions (e.g., interpersonal competence, emotional maturity, etc.). The ratings were completed independently by the supervisors of the combined sample of 243 participants. Results of correlations between the composite job performance measure and the three final in-basket factors revealed significant relationships for Organization of Work ($r = .20$) and Productivity ($r = .25$) but no significant result with Delegation was seen.

In sum, the Port Authority's landmark use of the in-basket exercise for personnel decisions was an historic step toward later widespread industrial applications of the instrument. The conclusions regarding the costly and complex nature of in-basket scoring from the 1960 police lieutenant promotion study clearly identified the scoring system as requiring further research and modification. Subsequently, the results of the in-basket exercises used in the 1964 management evaluation program provided limited evidence of criterion-related validity, although it did prove to be a very popular assessment device among candidates. Therefore, the in-basket exercise was retained as part of the management evaluation program and the scoring system was further refined in the following year's evaluation program.

After the 1965 application, Lopez (1966) optimistically concluded that the introduction of the Action Report form scoring method for in-basket evaluation was a promising way of assessing administrative ability and that it may represent a breakthrough in in-basket scoring. A careful consideration of the findings, however, suggests his optimism was not warranted. As noted in the Introduction, a key concern with the use of self-report questionnaires as the basis for in-basket scoring is whether candidates' actions from the exercise will be accurately reflected in the self-report. Either unintentionally, through carelessness, or intentionally, through choosing a more desirable course of action suggested by the questionnaire, the completed scoring form
may be distorted compared to candidates' actual performance. Lopez (1966) did little, empirically, to alleviate these concerns of motivational distortion and deception. The only information he provided about efforts to investigate such a serious disadvantage with self-report scoring was to assert that "later experimentation proved that these concerns were groundless, even in the competitive atmosphere of the assessment situation" (p. 109). This assertion remains questionable, because, to date, there appears to be no published empirical examinations of the extent or effects of deception or motivational distortion with the use of in-basket self-report forms. The self-report, questionnaire method of scoring was not widely adopted and did not represent the breakthrough Lopez had hoped. Nevertheless, Lopez continued to research the self-report scoring format and his additional findings will be reported shortly (Kesselman, Lopez & Lopez, 1982).

The Management Progress Study. The historical significance of this major longitudinal research of managerial assessment and career development by AT & T has been acknowledged in Section 1 of the Literature Review. Following the design principles of predictive validation previously outlined, the original assessment centre data gathered on 422 subjects were not released to company officials. Neither their performance nor their subsequent annual evaluations had any influence on the careers of the men being studied; hence, there was no contamination of subsequent criterion data by the assessment results.

The in-basket exercise used in the study was developed with the assistance of ETS. An interesting and unique deviation from standard ETS objective scoring principles (typically based on action elements) was that the evaluation of in-basket performance in the Management Progress Study was solely subjective, conducted by means of a 45 minute interview with the participant upon completion of the exercise. In the interview, questions were raised with participants concerning the approach they
took in the exercise, the reasons for specific actions taken, and their views of superiors, peers, and subordinates. After the interview, the rater prepared a detailed, written summary of in-basket performance. A set of guidelines in conducting the interview and a manual for report preparation were made available for raters to make the subjective evaluation process more reliable. It should be made clear that the in-basket exercise was not scored in any quantitative way during the assessment component of the study.

However, to determine the empirical relationship between in-basket performance and progress in management, a simple method to quantitatively evaluate the in-basket reports was derived. This post hoc method required two raters to independently read the performance reports and rate overall performance on a 5-point scale. By mutual agreement, a composite rating for each participant was determined. In 1966, ten years after the initiation of the program, the first set of predictive validity findings were published. At that time, data from 125 of the original 274 college-educated men and 144 of the original 148 non-college educated men were available. The criterion measure used in this analysis was salary progress, determined by taking the difference between salary at the time of assessment and that as of June 30, 1965. Across samples from seven telephone companies, in-basket validity coefficients ranged from -.19 to .44, with two coefficients reaching significance (.27 and .44).

Bray and Grant (1966) concluded that the two situational techniques (the in-basket and The Manufacturing Problem described on page 13), used in the Management Progress Study produced reasonably reliable results. In addition to evidence of predictive validity, both procedures were shown to have significantly influenced assessment staff evaluations (evidenced by correlations between situational test performance and assessment staff judgements). Bray and Grant were confident that neither technique could have been omitted without losing important information and
therefore, despite their high costs, continued use of situational exercises in assessment centres was justified.

**Cross' work with the school administration in-basket.** Furthering the research of Hemphill et al. (1962), Cross (1969) conducted an investigation into the predictive and concurrent validity of the Whitman School In-Basket used in the school administration study. With a sample of 14 school principals, the stability of stylistic in-basket scores and their relation to on-the-job measures of performance were examined by the administration of the first in-basket (labelled the predictive exercise), followed by a second administration of the same in-basket (labelled the concurrent exercise). Participants were given the predictive in-basket when they were involved in an administrator preparation program from 1961 to 1964 (not all subjects were involved in the program at the same time). In 1966, the second, concurrent in-basket was re-administered after they had been working in the school system for lengths of time ranging from 2 to 5 years across participants. This research design suggests a test-retest reliability paradigm allowing measurement of the stability of in-basket performance. However, because the time interval between administrations exceeded the six month maximum for assessing test-retest reliability advised by Anastasi (1982) and because the time interval was not consistent for all participants, it is not instructive to consider the correlations between predictive and concurrent in-basket scores.

Cross (1969) scored the in-basket responses using 26 of the 40 original stylistic categories identified by Hemphill et al. (1962)--the strategy for selecting the 26 categories was not reported. The empirical validity of the scorers' impressionistic ratings from the same ten adjective pairs used in Hemphill's earlier study were also re-examined by Cross. No attempt was made to score in-basket performance for content. The criterion measures were based on multiple observations of the principals at work, interviews with the principals, and analyses of samples of their written work. Separate
criterion measures were determined for each of the 26 stylistic in-basket categories. Little additional descriptive information on the measurement of the criterion variables was provided, except Cross' statement that "rules for scoring the stylistic measures of on-the-job behaviour were extrapolated from procedures for scoring stylistic in-basket performance" (p. 27). The reliabilities for these on-the-job stylistic behaviours were based exclusively on the observational data and were determined by correlating results across several observation periods. The reliability coefficients across the 26 categories ranged from -.04 to .87, and, in Cross' view, all but two were at an acceptable level.

For the predictive in-basket exercise, the validity coefficients ranged from -.45 to .59 with only two stylistic categories reaching significance (at the .05 level). Similarly, only two categories from the concurrent in-basket were significantly correlated with their corresponding measures of on-the-job behaviour, resulting in coefficients of .60 and .72. Scorers' impressionistic ratings of in-basket performance showed slightly more promising, but still weak evidence of empirical validity with 3 of 20 correlations (10 predictive and 10 concurrent coefficients) reaching significance.

Cross (1969) concluded that the in-basket exercise was best used as a training device instead of a device for predictive purposes. In considering whether the in-basket exercise could act as an effective selection device, Cross recognized that "the data do not permit a resounding affirmative response" (p. 28). It should be noted that the sample size in this study was exceedingly small and that conclusions regarding the lack of empirical validity for selection purposes using this instrument could only be tentative, at best.

IBM study. Wollowick and McNamara (1969) conducted a study of 94 male lower-to-middle managers from a large electronics firm (International Business Machines, or IBM) in order to determine the relative validities of the components of an
assessment center. The in-basket exercise was included as an individual, situational measure in the assessment program. The in-basket performance dimensions rated were Planning and Organizing, Self-Confidence, Decision-Making, Risk-Taking, Oral and Written Communication, and Administrative Ability, which were measured subjectively in an interview with an observer who inquired as to the actions taken by the candidate and the reasons for his decisions (no further scoring information was provided). The criterion measure used was the increase in managerial responsibility, measured by change in position level three years after the participants' initial involvement in the assessment program. Wollowick and McNamara reported a significant ($p < .01$) correlation coefficient of .32 between a composite in-basket score and increase in managerial responsibility.

The General Electric research. As with many earlier studies, Meyer's (1970) in-basket analyses at the General Electric Company were the result of collaboration between the company and ETS, in a specific effort to resolve selection problems for a particular middle management position. Following the in-basket design methods developed by Frederiksen et al. (1957) and Hemphill et al. (1962), approximately 50 stylistic categories of administrative performance were generated (e.g., Makes a Concluding Decision, Involves Subordinates, etc.). As discussed in the Reliability section, only 27 of the 50 stylistic categories were judged to have had adequate reliability to warrant retention in the study. The 27 categories were then factor analyzed and four in-basket factors emerged: Preparation for Decision, Taking Final Action, Organizing Systematically, and Orienting to Subordinate Needs. Finally, the Scorer's Rating—a subjective measure of overall performance based on the scorer's impression of the managerial skills shown in handling the items in the in-basket—was also examined as a predictor.
The criterion measures used in this study were generated from performance ratings of the 81 participating middle managers by their superiors on 12 "key areas of job responsibility" (these were not further described). A factor analysis was then conducted on these areas of job responsibility, yielding two job-performance factors: Supervision (concerned with human relations) and Planning-Administration (concerned with intellectual tasks). These two job-performance factors were then correlated with the scores from the 27 stylistic categories and with the four in-basket factors. At the five percent level, significant correlations between ratings of the Supervision factor and 3 of 27 performance categories were seen, ranging from .22 to .31, whereas the Planning-Administration factor yielded significant validity coefficients on 7 of 27 categories, ranging from .22 to .35. Two out of the four in-basket factors were significant when correlated against the Supervision factor ($r = .25$ and .32); similarly, two out of four in-basket factors were significantly correlated with the Planning-Administration factor ($r = .31$ and .40). Interestingly, the impressionistic Scorer's Rating yielded a greater, but nonsignificant, validity coefficient than two out of four in-basket performance factors when correlated against the Supervision factor ($r = .21$) and yielded a significant correlation greater than three out of four in-basket factors when correlated against the Planning-Administration factor ($r = .37$).

In addition to scoring for style, Meyer (1970) also measured the content of in-basket responses with an innovative and purely statistical, empirical approach rather than logical, panel-based approach. More specifically, he compared those courses of action taken by the sample of 81 managers who had received above average (median) on-the-job performance ratings on the two job-performance factors with those who had received below average performance ratings. Values were then assigned to each course of action based on its correlation with each of the two criterion measures (job-performance factors). No further information was provided regarding this unique
process, other than Meyer's statement that "weights were assigned and scores derived from courses of action in the same way as would be the case if empirical keys were developed for scoring item alternatives in a biographical inventory" (p. 302). In this way, two empirically-derived content keys for the courses of action were developed, one based on correlations with the Supervision factor and one based on correlations with the Planning-Administration factor. Validity coefficients between the scores derived from these two empirically-derived content keys and the two job-performance factors were not determined for the sample of 81 managers.

However, a second, smaller sample of 45 middle-level managers was administered the same in-basket exercise and content scores based on the two keys were calculated. Only the Planning-Administration content key showed evidence of criterion validity, with a significant correlation of .31 with the Planning-Administration job-performance factor. Additional correlations with scores based on the 27 individual stylistic categories were not calculated because, in Meyer's view, this second "cross-validation" sample was too small. Instead, 17 stylistic categories from two of the four in-basket factors (Preparation for Decision and Organizing Systematically) were combined to form a composite factor score, which was then correlated with the Planning-Administration job-performance factor in this second sample, resulting in a validity coefficient of .38 (significant at the .01 level). The Scorer's Rating was also re-correlated with the Planning-Administration job-performance factor; the validity coefficient increased from .37 in the first sample to .43 in this second sample.

Meyer (1970) concluded that the in-basket "might serve as a valuable aid in the selection of managers" (p. 306) and emphasized its unique advantage (compared to ability or aptitude tests) of high face validity. This distinctive feature is important, he claimed, because it made the exercise more acceptable to candidates and it resulted in straightforward interpretation which facilitated the feeding back of results to the
candidates. Meyer agreed with Lopez (1966) that a negative consequence of high face validity is that few studies of criterion-related validity of the in-basket have been conducted, despite its widespread use in managerial assessment. The results from Meyer's study indicated that certain styles in handling in-basket items correlate with on-the-job managerial performance. Despite these positive findings, Meyer called for additional research into the predictive validity of the in-basket, recognizing that, of the few in-basket studies conducted to date, most, including his, had been concurrent validation studies. Although Meyer did not emphasize the importance of his empirically-derived content scoring keys, this innovative scoring development forms the basis of one of three scoring strategies assessed in the present study.

A "leadership" in-basket exercise. Brass and Oldham (1976), like Lopez (1966) and Meyer (1970), cited the paucity of in-basket criterion validation research. Brass and Oldham further claimed that the few findings which had been reported had not adequately established whether in-basket performance corresponded to actual managerial performance; "few of the correlations between scores on the various in-basket scoring categories and managers' on-the-job performance ratings have reached acceptable levels of significance" (p. 652). They speculated that a chief reason for these disappointing validation results may have been that in-basket scoring categories used by researchers did not accurately reflect effective managerial behaviour. According to Brass and Oldham, scoring well on in-basket performance dimensions bore little or no relation to effectively performing one's duties as a manager.

The target, therefore, for their research efforts was the derivation of the dimensions or categories used to score the in-basket that would more accurately reflect effective managerial behaviour. Brass and Oldham's (1976) approach was to score participants on behavioural dimensions that have been shown to reliably predict managerial and subordinate performance. Specifically, they identified leadership
ability as a critical measure of managerial effectiveness. The basis for the derivation of the in-basket dimensions was a theory of "environmental control" involving the degree to which the manager enhanced the work motivation and performance of subordinates by controlling their work environments (Oldham, 1976). Accordingly, by operationalizing aspects of leadership based on this notion of environmental control, six specific leadership activities were selected as in-basket categories: Personally Punishing, Personally Rewarding, Setting Goals, Placing Personnel, Designing Job Systems, and Designing Feedback Systems.

A sample of 71 male first-level foremen from a large American manufacturing company were administered this leadership in-basket exercise designed specifically for this study. A scoring manual was developed which set out guidelines and specific rules for scoring each of the six leadership activities in relation to the 28 items comprising the in-basket. Following ETS principles of item preparation, each of the items was designed to provide an opportunity for one or more of the leadership activities to be demonstrated. Scorers examined the response(s) to each item and assigned a 1 or 0 to reflect whether or not each of the particular six leadership categories had been evidenced by the participant. Multiple occurrences of a particular leadership activity within one item were not scored and so the maximum score for a single leadership category was 28. No other aspects of in-basket performance (i.e., scorers' impressionistic ratings, and content, as opposed to style) were measured.

The criterion data used by Brass and Oldham (1976) were supervisory ratings provided by the immediate supervisors of the foremen. By adapting procedures suggested by Smith and Kendall (1963), eight measures of managerial effectiveness were derived, each based on critical incidents of both effective and ineffective foremen behaviour. The supervisors were asked to sort the incidents into their appropriate performance categories and scale them on a 7-point effectiveness continuum. A single
composite measure of managerial effectiveness was then derived, based on an average of the eight measures. Reliability results for the criteria were not reported.

When correlated against the composite criterion measure, four of the six leadership categories were positively and significantly related, with validity coefficients of .24 (p < .05), .27, .28, and .34 (p's < .01). Foremen who personally rewarded their subordinates for good work, who punished subordinates for poor work, who set specific performance objectives, or who enriched their subordinates' jobs also tended to be rated by their supervisors as effective performers. Brass and Oldham (1976) recognized that, because some of the leadership categories were used infrequently, some of the validity coefficients may have been inflated (frequency data were not provided). For this reason, they advised caution in the interpretation of the validity coefficients.

Nevertheless, Brass and Oldham (1976) firmly concluded that the results of their investigation of the criterion-related validity of the in-basket were substantially stronger than past findings. They asserted that such strong findings were due, in large part, to their focus on the derivation and selection of in-basket scoring categories which were relevant to actual managerial performance. It may also be that using a narrowband instrument (leadership ability is only one of several components of general managerial behaviour) allows a tighter, more controlled operationalization of the scoring categories, making measurement more objective and accurate. However, the relatively restricted measurement of managerial behaviour used by Brass and Oldham's leadership in-basket may preclude the generalizability of their positive findings to situations where managerial performance is more broadly-defined (as is typically the case).
A return to the self-report scored in-basket exercise. Continuing the novel approach reported by Lopez (1966), Kesselman et al. (1982) designed an in-basket exercise scoring format based on a self-report questionnaire (the Action Report form). A group of 85 first-line supervisors (65 male and 20 female) from a utility company, upon completion of the exercise, were provided with a list of all possible actions for each of the 26 items included in this exercise. The list was constructed by first grouping possible actions globally (e.g., "I communicated with Shivers as follows:") and then providing more specific options within that grouping (e.g., 

#171. Contact Riddle, determine priority and follow-up on request; #172. Have Riddle send me his performance data for review"). There were, in total, 684 possible actions listed. Respondents completed the Action Report form by choosing the appropriate numbered option(s) that corresponded to the action(s) they took. In addition, they were asked to rate the priority of each item and provide the reasons for the action taken. The Action Report form included an "Unusual Action" form wherein candidates could add actions that were not already listed. Kesselman et al. reported that only 4% of candidates used the Unusual action form, and nearly all of the unusual responses could be classified into one of the 684 listed actions. A total of three hours was required for administration with two hours needed to complete the in-basket exercise and one hour to fill in the scoring report form.

The participants' supervisory positions involved two job areas, classified as administrative and physical/technical. Standard job analysis techniques (i.e., interviews, questionnaires, work samples) were employed in both job areas to determine the in-basket performance dimensions to be measured and to provide information for in-basket item preparation. In addition, a method called Threshold Traits Analysis (TTA) was used. TTA lists 33 traits thought to be necessary for effective job performance and requires supervisors of the job being analyzed to
determine the relevance, level, and practicality of each trait (Lopez, Kesselman & Lopez, 1981). The results from both the TTA and standard job analysis techniques provided a list of specific traits required for effective job performance across both job areas. Three traits were selected and the following three in-basket dimensions, based on the three traits, were developed: (a) Problem-Solving, (b) Planning, and (c) Decision-Making. Because of high intercorrelations among these dimensions, scores were subsequently summed to yield a total in-basket score.

Kesselman et al. (1982) selected ten middle managers to form a panel of subject matter experts who were asked to assess the appropriateness of each of the 684 possible actions and assign scoring weights, by group consensus, based on a scale from 0 (inappropriate) to 3 (highly appropriate) for each of the three in-basket dimensions. In addition, each panel member was asked to rate each action on a 1 to 4 scale reflecting the priority of the item (immediacy of action required). Final scoring weights for each possible action were then determined by multiplying the appropriateness value by the priority rating.

Next, to establish the job performance criteria, eight performance dimensions, based on two separate measures, were completed by the immediate supervisors of the participants. The two separate measures both identified a set of traits related to job performance, but used two different measurement methods. The first performance measure used a graphic scale to measure job performance on each of the 33 traits based exclusively on the TTA (no further details were provided). The second performance measure included traits from the TTA as well as traits derived from standard job analysis techniques. Two 6-point behavioural observation scales were used to measure job performance based on the set of combined traits, one for the administrative and one for the physical/technical positions. The two types performance measures (also called forms) were correlated, and an average "inter-form" reliability of .70 was reported.
No further criterion reliability information was provided. Ultimately, four of the eight job performance dimensions were chosen from each of the first and second measures. In addition, an Overall Job Performance rating for each of the two measures was determined. The results using this criterion measure (only) will now be related.

The validity coefficients reported by Kesselman et al. (1982), based on correlations between the total in-basket score (summed across the three in-basket dimensions) and the Overall Job Performance rating from the two forms, were .33 and .27, both significant at the .01 level. Corrected coefficients, adjusted for attenuation in the criteria, were also reported, based (apparently) on the earlier inter-form reliability estimate of .70. The corrected correlations were .38 and .31. The individual validity coefficients for the three in-basket dimensions were not reported.

In sum, Kesselman et al. (1982) sought to clarify what they believed were mixed criterion-related validity results reported for situational exercises. In their view, the limited research evidence did not unequivocally support the use of situational exercises over paper-and-pencil tests. At the same time, they also sought to develop a scoring format which was less costly and less time-consuming than other in-basket scoring procedures. Kesselman et al. concluded that, with their refinements to and application of the Action Report form, the results showed that their in-basket scoring method reliably, accurately and economically predicted managerial ability. However, the same problems of deception and motivational distortion which plagued Lopez’s (1966) self-report form remained unaddressed. Kesselman et al. acknowledged that future research should be directed to address both the extent and impact of deception in the self-report format. As noted earlier, despite continued development and application of self-report scored in-baskets, there appears to be little or no published empirical examinations of the extent or effects of deception or motivational distortion. In addition, the use of group consensus in this study to assign scoring weights to the
courses of action is a questionable methodological practice because of the likely effects of undesirable group dynamics in what should be an objective, empirical process. According to ETS, factors like the undue influence from one or two "strong personalities" within the group, "political" pressures from using members of different hierarchical levels within the company, and physical fatigue leading to a "let's just get this done" approach are some examples of the serious, subjective flaws inherent with the process of group consensus to determine scoring weights (K. Abbey, personal communication, May 8, 1991). Consequently, ETS typically solicits individual, independent (as opposed to group-based, inter-dependent) ratings from panel members to determine scoring weights for the appropriateness of the courses of action.

A predictive validation study. A large-scale study by Turnage and Muchinsky (1984) examined the predictive validity of assessment centre evaluations and several "traditional" predictor variables in predicting managerial performance. The in-basket exercise was one of a battery of assessment devices administered, over a four year period, to a sample of 799 employees (92% male) in a large manufacturing firm. No details were provided regarding the design or scoring of the in-basket; it is not known whether a purely subjective (i.e., interview) or a combination (interview and objective) scoring method was used. Ten criterion variables were used, including salary progress, promotions, career potential ratings, and standard job performance ratings. As with the predictor variables, little specific information (such as development, descriptions, reliability, etc.) was provided regarding the criterion variables. None of the traditional predictor variables, including the in-basket exercise, was found to be predictive of salary progression, job performance ratings, or promotions. In-basket performance, however, was found to be significantly related (at the .01 level) to ratings of career potential ($r = .25$).
A quickly-scored in-basket. Hakstian, Woolsey and Schroeder (1986) developed an in-basket exercise for the prediction of first-level supervisory performance which could be scored quickly by hourly workers in a provincial telephone company. From an initial pool of six in-basket performance dimensions, Productivity, measured by the number of items dealt with and the approximate number of words per item written (summed over all items), and Content, measured by an evaluation of 10 of 22 items administered as part of the exercise, showed evidence of concurrent validity. The Content scoring key was generated by a three-member panel who examined each of the ten items and derived a series of 2-point, 1-point and 0-point responses. Unanimity among the panel members was needed to make final Content value assignments.

The criterion measures were provided by a 5-point behaviourally-anchored performance appraisal which measured seven performance dimensions and was completed by the immediate supervisors of the 238 participants (128 men and 110 women). Two global criterion measures, Overall Work Performance and Overall Interpersonal Skills, were generated by taking the first principal component of scores for the appropriate section of the performance appraisal. When correlated with Overall Work Performance, the in-basket Productivity dimension yielded validity coefficients, significant at the .01 level, of .23 (for males) and .26 (females). With this same criterion measure, the in-basket Content category was not significantly related to performance with either gender. Using Overall Interpersonal Skills as the criterion, only the Productivity dimension applied in the female subsample resulted in a significant correlation ($r = .25$).

Summary of criterion-related validity evidence. The in-basket research reviewed here may appear to present substantial evidence of criterion-related validity. However, a closer inspection of the data reveals a collection of disparate instruments applied in a variety of settings with little information provided on methodology. Many
of the research settings are not industrial (e.g., military, academic, etc.), and there are
great variations in the constructs measured and outcome criteria used. Also, there is
little consistency in the design format and scoring approaches of the in-baskets
reviewed. Given the disparity in test construction, methodology, performance
dimensions, and outcome criteria, perhaps it is not surprising that reviews of in-basket
criterion-related validity results appear mixed (Gill, 1979; Schippmann et al., 1990;
Thornton & Byham, 1982).

Notwithstanding the disparate methods and settings of the in-basket studies
reviewed, an approximate average range of the validity coefficients reported in Table 1
is between .25 and .30. However, it should be recognized that predictive validities for
the in-basket in the high .20's and low .30's are not inconsequential. In a utility
analysis of the assessment centre for selection, Cascio and Silbey (1978) reported that
even with a validity of .05, the gain in the dollar-value utility of the assessment centre
over random selection represented a net savings of $173 per selectee. The in-basket
exercise, which makes a significant and unique contribution to the overall assessment
centre rating (Huck, 1974), no doubt results in substantial dollar-value utility gains
when used in the assessment centre as a selection device.

Construct Validity

School administration in-baskets. In addition to gathering evidence for
criterion-related validity, a key purpose of the school administration study (Hemphill et
al., 1962) was to investigate the relationships between the major dimensions of
administrative performance and a variety of other personal characteristics. A series of
cognitive ability tests, a personality inventory, and an interest inventory were among
the measures used to gather data for a broad set of characteristics. Specifically, the
Strong Vocational Interest Blank for Men (Strong, 1951), the Sixteen Personality
Factor Questionnaire (Cattell, 1957), four fluency factors (Guilford, 1957), and a set of ability tests from the Kit of Selected Tests for Reference Aptitude and Achievement Factors (French, 1951) were employed. It is not appropriate to summarize all findings here, except to briefly describe the strongest relationships between the most salient stylistic factors of in-basket performance and the set of personal correlates examined in the study.

Those school principals who scored high on the in-basket Exchanging Information dimension also tended to have a strong vocabulary and be emotionally sensitive. Strong skills in oral communication were associated with high scores on Analyzing the Situation and Organizing Work. The tendency to organize work as measured by the in-basket, in turn, was positively related to principals who were anxious, insecure and characterized by nervous tension. Strong skills in word and ideational fluency and skill with simple numerical tasks were characteristic of principals who scored high on Complying with Suggestions made by Others dimension but low on Directing the Work of Others. Principals who were friendly and adventurous were more likely to be involved in maintaining relationships with others in the school portrayed in the exercise. In general, most of the in-basket stylistic factors were unrelated to interest scores.

Frederiksen (1966). A similar study was later conducted wherein in-basket performance from a sample of 115 male administrators in the federal government was related to ability, personality, and interest factors. Among his findings, Frederiksen reported that those who tended to take many imaginative courses of action also tended to have higher scores on a vocabulary test. Also, those who wrote a great deal, attempted many items, involved many people, and took many leading actions also tended to have high scores on the Active scale of the Thurstone Temperament Schedule (Thurstone, 1953). Those who planned to have many face-to-face discussions with
subordinates tended to resemble life insurance salesmen in their responses to the Strong Vocational Interest Blank, whereas those who resembled forest service men tended to avoid discussions and asking for information.

A parallel form study. Brannick et al. (1989) conducted an examination of two parallel forms of an in-basket exercise, developed as part of a statewide managerial training program, in order to study the construct validity of in-basket scores. Using a pretest-posttest design, 88 university students (49 males, 39 females) were first randomly assigned to take either Form A of Form B of the in-basket (pretest). One month later, the students then took the in-basket that was the alternate form of the one taken previously (posttest). The in-baskets forms were designed, following standard ETS principles, to measure Organization, Leadership, Perceptiveness, Decision-making, and Delegating. Separate scoring keys were developed for each in-basket after first constructing a list of the usual courses of action for each item. For each form, and for each item within that form, each response was judged as positive, neutral, or negative and assigned a +1, 0 or -1, on each of the five dimensions. It should be noted that this application of qualitative or evaluative-based scoring of dimensions is a departure from the standard ETS scoring practice of using quantitative or frequency-based scoring. The final keying of each response was determined by consensus of a panel of five (three authors and two scorers). A total score was calculated by summing the dimension scores.

A multitrait-multimethod matrix (Campbell & Fiske, 1959) was set up to examine the convergent and discriminant validity of in-basket performance. According to Campbell (1960), in order to demonstrate construct validity, a test should be shown to both correlate highly with other variables with which it should theoretically correlate (convergent validity) and to not correlate significantly with variables from which it should differ (discriminant validity). Therefore, Brannick et al. (1989) divided the two
parallel in-basket forms into two halves, creating four distinct exercises. One half of each exercise contained scores on the even numbered items, and the other half contained scores on the odd numbered items. Generally, the validity diagonal results were non-zero (Form A: r's ranged from .19 to .61; Form B: r's ranged from .20 to .62), but, for the most part, they were not larger than the off-diagonal values.

Brannick et al. (1989) concluded that, although there was limited evidence of convergent validity and even less evidence of discriminant validity. In their view, the data "...failed to support the interpretation of in-basket scores as indicants of managerial ability" (p. 962). This conclusion, while distressing, may be explained, in part, by the scoring system employed in the study. In addition to scoring for actions taken, candidates were also scored for actions not taken. Failing to take action on an urgent item was typically seen as a negative response and, because each action was evaluated along each dimension, such non-action was usually scored negatively across several (if not all) dimensions. As the researchers themselves acknowledged, scoring the same behaviour, in the same way, for several managerial abilities would likely result in high correlations among dimensions. It is not surprising, then, that little evidence of discriminant validity was seen; it is unlikely that multiple dimensions, measured in this way, would satisfy the "heterotrait" condition of multitrait-multimethod design.

Similarly, if the methods used in a multitrait-multimethod matrix are not substantially different, it is unlikely that a true "heteromethod" situation exists. The absence of true "heteromethod" comparisons would likely result in higher off-diagonal correlations. It is questionable whether it is reasonable to expect evidence of discriminant validity when alternate forms of the same instrument are used. Although the assessment of administrative ability does not easily lend itself to multiple methods of measurement, it may be more instructive to conduct convergent and discriminant
validation using in-baskets which differ in the nature of the stimuli (e.g., paper-and-pencil versus videotaped presentation), and the nature of the response format (e.g., open-ended versus multiple-choice), in order to more closely replicate a multitrait-multimethod design, and so more soundly demonstrate construct validity (or the lack of).

The measurement of participative decision-making (PDM). Tett and Jackson (1990) designed and evaluated an in-basket measure of managerial participative decision-making and examined the relations between participative tendency and various personality traits. The identification of in-basket dimensions was theory-driven (Jago, 1978; Vroom & Yetton, 1973) and resulted in six participative behaviours to be measured in the exercise: Delegation of Decision-Making Authority, Requesting Advice, Following Advice, Requesting to Meet with Subordinates to Discuss a Problem, Requesting Information, and Asking to be Kept Informed (as to how a problem is developing or being resolved). Items were prepared, using standard job-analysis techniques, to be realistic and representative of typical managerial problems. The process of deciding which dimensions were relevant to which items was not clarified.

The scoring system for the six participative dimensions was unique, with the number of subordinates involved forming the score for the dimensions Requesting to Meet, Requesting Information, Requesting Advice, and Asking to be Kept Informed. The Delegation dimension was scored on a scale from 1 (explicit directions) to 6 (complete delegation). Following Advice was scored as 0 for advice rejection, 1 for a neutral response, and 2 for advice acceptance. Scores within a particular dimension and across items were averaged to yield six participative dimension scores. Dimension scores were then summed to provide an overall PDM score.
Three theory-based components of participative decision-making (power sharing, interactions among co-workers, and information exchange) related to worker satisfaction (Wall & Lischeron, 1977) were used as a framework to then select a total of six related personality traits from the Personality Research Form (Jackson, 1986) and the Jackson Personality Inventory (Jackson, 1976). The component of power sharing led to the selection of Dominance and Autonomy as personality correlates in this study; interactions among co-workers provided Affiliation and Social Recognition; information exchange pointed to Cognitive Structure and Tolerance. (External criterion validation of the in-basket exercise was not carried out in this study.)

The sample consisted of 89 mid- to upper-level managers (82 men, 7 women). When the personality traits were correlated with the overall PDM score, three of the six personality traits were found to be unrelated to participative tendency and two were correlated in the direction opposite to that predicted. These unexpected positive correlations were between Delegation and the traits of Autonomy and Dominance ($r = .23$ and $r = .36$, respectively, both significant at the .05 level), and lead Tett and Jackson (1990) to recognize the need for further work into the effect of authority-threatening situations in the relation between power-based motives and participative decision-making. In addition, following a principal components analysis, evidence of the unidimensionality of the participative behaviours was found. Consistent with the trend toward computer scoring (Thornton, 1992), a computerized version of the PDM in-basket exercise is under development, based on a multiple-choice response format, to "...permit highly reliable but less labor-intensive scoring" (Tett & Jackson, p. 181).

Another multitrait-multimethod study. Recently, dimensions from the in-basket featured in the present study (Telephone Supervisor In-Basket Exercise, or TSIB) and the most closely-corresponding dimensions from the ETS Consolidated Fund In-Basket Test (Educational Testing Service, c. 1970) were combined to form a multitrait-
multimethod matrix (Hakstian & Harlos, 1992). Because both instruments will later be discussed in detail, only the results of the study will be related here. (A limited assessment of the construct validity of the TSIB will also be reported under Study 2 of the Preliminary Studies section.)

Consistent with the findings from Brannick et al. (1989), some evidence for convergent, but little for discriminant, validity was seen. The mean validity diagonal was .43 (without disattenuation), and although the mean heterotrait-heteromethod correlation was lower (.32), three of the five TSIB dimensions correlated more highly with non-corresponding than with corresponding ETS in-basket dimensions. Unlike the in-baskets used by Brannick et al. (1989), the TSIB used more standard ETS scoring protocols. However, like Brannick et al. (1989), multiple keying of courses of action across dimensions was common in the TSIB, creating dimensional dependency. Perhaps a multitrait-multimethod matrix based on factor-analytically derived in-basket dimensions would be instructive, reducing the contribution to variance from dimensional dependency and allowing a more focused analysis of discriminant validity.

Factor-analytic studies. Factor-analytic studies of in-basket performance may provide more indirect evidence of construct validity than those studies relating in-basket performance to data from external (i.e., non-test) variables (Schippmann et al., 1990). Nevertheless, these findings merit consideration because they do provide additional information about the nature of in-basket performance. Factor-analytic studies of the correlations of large numbers of in-basket style categories have suggested some common dimensions of administrative performance: Complying with Suggestions, Preparing for Action by Gathering More Information, Directing Others, and Discussing Problems with Others (Thornton & Byham, 1982). The factor consistency across studies (and high intercorrelations among dimensions in non-factor-analytic studies) has led some researchers to postulate that the underlying characteristics measured by the in-
basket exercise are based on a single generalized trait (Kesselman et al., 1982; Lopez, 1966). Other researchers, however, have contended that the factor consistency ".. may represent nothing more than the potentially clouding influence of an ETS dominated measurement system" (Schippmann et al., 1990, p. 856).

**Summary of construct-related validity evidence.** Considering the widespread use of in-basket exercises, there are few studies reported in the literature which directly assess construct validity. The limited evidence suggests that, particularly when related to cognitive ability, in-basket performance reflects logical and generally consistent characteristics. The in-basket exercise appears capable of measuring specific theoretical constructs. However, factor-analytic findings have resulted in some disagreement among researchers as to the complexity of administrative performance. On one hand, it is believed that skills in this area are multi-dimensional, that several separate, identifiable styles and action-approaches are present and thus need to be measured to fully understand and to make accurate predictions about administrative potential (Frederiksen, 1966; Thornton & Byham, 1982). Conversely, some researchers maintain that a more global, unitary skill underlies performance (Kesselman et al., 1982; Lopez, 1966).

*Content Validity*

In general, the realism, relevance, and representativeness of items in an in-basket contribute to content validity. As noted earlier, conducting thorough job analyses in order to more realistically reflect the nature and complexity of the target job substantially enhances content validity. Although the theoretical procedures to *establish* content validity for in-baskets are clear, in the view of Schippmann et al. (1990), "all of the reviewed studies which suggest that their procedures are [italics added] content valid fall seriously short of the mark" (p. 851). The reason the studies fell short, they
contended, is that researchers have confused face validity with content validity; the exercise, on the face of it, may seem to reflect the challenges present in a position, and so proper content validation is not completed. Another factor which may discourage test developers from adequately establishing content validity is the length of time required to sample and integrate documentation from real in-baskets (Gill, 1979). Schippmann et al. concluded their review of content validity findings by recognizing that "there simply are no published or widely distributed reports which describe how to develop an in-basket that is well-grounded with regard to content validity" (p.851).

General Summary of In-Basket Validity Results

Despite the range of in-basket scoring approaches developed and researched, it is simply not possible to know, with confidence, whether any one key is correct. For several reasons, accurate assessment of in-basket validity is a daunting task. Because of its multi-faceted nature and the lack of clear, consistent definitions and descriptions of managerial behaviour, there are inherent limitations in accessing and measuring managerial performance. In addition, problems of judgement bias remain a serious obstacle in performance appraisal assessment (Cascio, 1987).

As previously noted, it is difficult to summarize published empirical findings of the in-basket because important detail is often missing in the reporting of methods, scoring procedures, and results. Moreover, substantial variations in research settings, in-basket construction and scoring, performance dimensions, level of exercise fidelity, and external criterion measures make meaningful comparisons difficult. Despite these difficulties, some summary observations can be made which point to several areas in need of further in-basket research.

The equivocal support among reviewers regarding the use of the in-basket exercise as a selection measure suggests that further evidence of criterion-related
validity is called for. Several explanations have been proposed for the relative lack of consistent, sound empirical validation for this instrument, including its high face validity, lengthy time requirements to train in-house scorers, and the difficulty of scoring the exercise, and the distortion that arises from subjective evaluations of in-basket performance. Industry users may believe that further research is unnecessary, either due to high face validity or due to the inherent soundness suggested by the lengthy training of scorers and the complexity of scoring (Lopez, 1966; Kesselman et al., 1982; Thornton & Byham, 1982). Even when in-basket research is undertaken, sound results may be hard to obtain because of the difficulty in properly constructing in-basket exercises and the reliance in assessment center applications of the in-basket exercise on subjective evaluation methods. Regardless of the reasons, more evidence of criterion-related validity appears warranted before it can be concluded, with confidence, that the in-basket exercise possesses the predictive properties to justify its widespread use in industry.

Further evidence of construct validity also appears warranted, because of both the lack of published studies directly assessing the nature of in-basket performance and the controversy surrounding those results which have been reported. Thus, additional research into the nature and number of in-basket dimensions that are involved in administrative performance is needed. Finally, there is a need for clear guidelines to develop content-valid in-basket exercises which must be consistently applied across industry.

This review of the content-, construct-, and criterion-related evidence of validation suggests that, for the most part, Frederiksen's (1957) goal of designing an instrument which is "sensitive" (can adequately measure the broad, complex set of skills involved in high-level jobs) has not yet been satisfactorily met. This is especially apparent when one considers the overwhelming gender imbalance of subjects in studies
assessing criterion-related validity. Over the forty years of in-basket research in those studies which clearly identified gender, roughly 1,672 men, compared to 290 women, were involved as subjects. (The same trend, although to a lesser degree, is apparent in the few published studies of construct validity.) It is assumed that, in studies where the gender of subjects was not clearly specified (e.g., Lopez, 1966), subjects were male. The reasons for such a disturbing discrepancy are not relevant here. What is relevant is the realization that the findings and summary observations from published in-basket research cannot be assumed, in all cases, to apply to women. With gender-based differences in cognitive ability it is not unreasonable, at the least, to gather additional data on women's in-basket performance, thereby setting the groundwork to further investigate the possibility of differential attributes and predictive accuracy of administrative ability based on gender.

**Summary of Section 2: Psychometric Properties of the In-Basket Exercise**

The second section of the Literature Review considered, in some detail, the pioneering work by Frederiksen in the early 1950's, which laid the foundation for conventional in-basket design and traditional objective scoring strategies. A chronological review of the published empirical findings of the in-basket exercise was then presented, which included both a description of design and procedural developments, as well as a summary of the reported psychometric properties of the instrument. Several significant inadequacies in both the quantity and quality of the research were identified and areas for further research were discussed.
Section 3: A Review of Current In-Basket Scoring Strategies

Limitations of a Literature-Based Review

In contrast to the disparate studies and contradictory findings reported in Section 2 (Psychometric Properties), the difficulty faced by researchers in scoring in-baskets is a uniquely consistent observation. Since Frederiksen et al.'s (1957) disappointing psychometric results, research which considers both psychometric and practical implications of scoring methods has continued, albeit slowly, resulting in the development of several scoring approaches. The Introduction described these three main approaches of evaluating in-basket performance: a) qualitative, subjective, clinical (typically based on a post-exercise interview); b) quantitative, objective, psychometric (based on analyses of written responses); and c) a combination of the two. The psychometric approach was further broken down into two methods: the more traditional, ETS-based scoring method based on molecular courses of action (also known as action elements), and the less conventional, self-report multiple-choice form.

It should be noted that the classification of scoring strategies introduced here is not present in the literature. Brief reviews of general scoring practices have appeared (Gill, 1979; Kesselman et al., 1982), but these tend to be too cursory to be useful. Currently, neither a systematic delineation nor a systematic evaluation of scoring approaches exists. As acknowledged earlier, many studies fail to provide adequate detail on the scoring method used, and, as a result, valid comparisons across studies are often precluded. In one of the few references to the combination method of scoring in-basket performance, Clutterbuck (1974) reported one industry user's claim that the combined interview and psychometric approach could increase the efficacy of the in-basket by as much as 50 per cent. This assertion, however, is purely subjective. To date, no known studies examining the combination method of in-basket scoring have
been published. In one of two major reviews of in-basket research, Gill's (1979) review did not even refer to the ETS action element approach in his description of the quantitative scoring method, instead selecting only the multiple-choice self-report format.

With the widely-acknowledged difficulty of in-basket scoring and without a systematic, comprehensive delineation of scoring approaches, it is not surprising that very few studies have focused specifically on scoring methods. Consequently, research and opinion into the relative merits (psychometrically and practically) of the various scoring strategies are inconclusive (Gill, 1979). As noted in the Introduction, Brannick et al. (1989) cited comparisons of in-basket scoring systems as a principal area for much-needed investigation.

Given the limited findings on in-basket scoring, inaccuracies in the few published reviews of in-basket research make the accurate assessment of in-basket scoring even more difficult. For example, Gill (1979) falsely described Meyer's (1970) analysis of in-basket performance as resulting in two in-basket performance dimensions for scoring, namely, a supervision and a planning/administrative dimension. As discussed in Section 2 of the Literature Review, these dimensions were job-performance dimensions, not in-basket performance dimensions. Gill compounded his error by asserting that many other studies supported Meyer's finding of "in-basket" dimensions of supervision and planning/administration to be used for scoring in-basket performance. In another misstatement, Gill claimed that the study by Wollowick and McNamara (1969) was one of the few studies which showed that the objective, psychometric approach to in-basket scoring was superior. Yet, as Section 2 indicated, Wollowick and McNamara's very limited discussion of the in-basket scoring method they used clearly identified it as the subjective (interview) approach. They had reported that a subjectively-derived Overall Assessment Rating (OAR) yielded
substantially lower validities in the prediction of management success than an empirically-derived combination of scores and ratings obtained across performance measures. Moreover, Wollowick and McNamara's conclusion of the superiority of the statistical (objective) scoring approach was, in fact, relevant only to methods used to combine variables in the derivation of the OAR, not in-basket scoring. Because of the importance of Gill's work as one of only two reviews of in-basket research, the misidentification of in-basket scoring dimensions and the inaccuracy regarding the relative efficacy of scoring methods may, if unrecognized as erroneous, make the synthesis of findings (which already are disparate and insufficiently detailed) even less conclusive.

Considering the paucity of sound research (or reviews) conducted on in-basket scoring, it may be more instructive to turn to current industry practice, rather than rely on the literature, in order to gather more complete and accurate information of current in-basket scoring strategies.

Review of Current In-Basket Scoring Methodology in Industry

Serious obstacles exist which make the goal of synthesizing industry-based scoring information laborious, if not unattainable. Detailed descriptions of the scoring protocols and specific concepts used in the development and application of industry in-baskets are simply not available. Many companies, through consultants, opt for custom-designed, "in-house" in-baskets, rather than "off-the-shelf" exercises. The concerns from developers of both in-house and off-the-shelf in-basket exercises regarding competition from other in-basket consulting rivals have led to a fiercely protective, vigilant control over "design and scoring secrets." Despite the difficulties in surveying in-basket scoring methodology used in industry, some important information regarding several widely-used in-basket exercises has been ascertained. What follows is a brief review of the design and scoring of these instruments.
The Multiple-Choice In-Basket Management Exercise (MCIME). Morris (1991) has developed a wideband, low-fidelity simulation, very similar in format and scoring to that described by Motowidlo et al. (1990). According to Morris, the construction of the MCIME was based, in part, on a request from the U. S. Equal Employment Opportunity Commission for alternatives to traditional multiple-choice tests used in selection. In addition, the evidence that assessment centres are good predictors of managerial potential (Cohen, Moses & Byham, 1974), that assessment centres do not result in adverse impact for minorities or women (Huck & Bray, 1976), that in-baskets correlate well with OAR's (Huck, 1974), and that in-baskets themselves are good predictors of managerial performance, also led Morris to consider the application of a multiple-choice format in-basket to the assessment centre method.

The MCIME consists of 81 questions or items presented to the participant with four response options provided for each question. The participant is directed to select the option representing the "best" and "worst" way of handling each problem using a strictly multiple-choice format; no free-format written responses are made. Standard job analysis techniques were used to both prepare items and identify dimensions relevant to performance as a manager. The in-basket items were ultimately prepared to measure six performance dimensions: Planning and Organizing, Decision-Making, Interpersonal Relations, Problem Analysis and Issue Identification, Written Communication, and Overall Job Performance. The MCIME is unique in that no time limit is imposed for completing the exercise.

A racially and gender-balanced panel of subject matter experts used the consensus approach to determine the scoring key for the in-basket dimensions, first by deciding which dimensions apply to each item and then by reaching a consensus on the best and worst options for handling each item. One point each (toward the dimension score) is assigned for the participant's correct (as defined by the panel) selection of best
and worst option. Thus, for each item, scores range from 0 to 2, with 2 reflecting the accurate selection of both the best and worst options. Multiple keying of items to dimensions does occur, from a minimum of 13 items used to measure Written Communication to a maximum of 49 items being scored for Problem Analysis and Issue Identification. Feedback is in the form of a computer-generated diagnostic report prepared for each participant.

The objective scoring approach used in the MCIME represents a recent revision to a previous objective scoring approach with the result that little empirical data are available. Nevertheless, a recent study of 150 subjects using the newly-revised objective scoring approach obtained uncorrected validity coefficients ranging from .26 to .42 across the six dimensions measured (D. Morris, personal communication, October, 1991). Additional information about either the MCIME itself or related empirical findings are simply not attainable. Efforts are currently underway to prepare and publish a synopsis of the design, development, and predictive accuracy of the instrument. For this reason, until publication, Morris wishes to restrict the release of information through other sources (D. Morris, personal communication, June 29, 1992).

**General Management In-Basket (GMIB).** Another objective scoring method is applied by Joines (1991), who has developed a more traditional, medium-fidelity in-basket exercise wherein participants read a series of realistic-looking items and, on standardized forms, provide both an analysis of the managerial issues involved in each item and a description of the actions they would take in handling the item. In addition, on a second standardized form, participants write out responses or memos to fully execute their actions in completing the item.
The GMIB's scoring principles reflect its intended development as a generic in-basket, designed to measure managerial skills regardless of the target job classification. Its content is described as theory-driven, in that many items are written to require participants to apply sound management theory to practice. Management concepts from McGregor's Theory Y (1960), participative management, and situational leadership form the basis not only of the instrument's content, but also of its scoring system. Unlike the standard ETS panel approach, the GMIB scoring keys reflect the views of the test developer, not industry personnel. The scoring key is based more on acceptance and application of the relevant management theory for each item (as defined by the test developer) rather than on the consensus judgements of subject matter experts or panels.

The GMIB also differs from ETS-developed in-basket exercises in that no Reasons for Action form is involved. In-basket evaluation is based solely on the participants' written responses to the 15 items that make up the GMIB. Three of the 15 items are considered critical and these are scored on a scale from 0 to 5 whereas the remaining 12 items are scored using values ranging from 0 to 4. Scoring values for courses of action are assigned by the test developers who have also provided response examples, set out in a scoring manual, for each scoring value. According to Joines (1991), trained raters require only 20 minutes to score one GMIB, including selecting, for each item, an appropriate narrative statement from a bank of statements. Using this approach, a 5-6 page narrative report is prepared for each participant.

Joines (1991) reported that four reliability studies have been conducted on the GMIB. The largest of these used two raters who each scored 100 exercises. The inter-rater reliability coefficient, based on the total in-basket score, was .95. In addition, an alpha coefficient of .71 was reported for the total GMIB score.
A large-scale criterion-validation study of the GMIB was conducted with a sample of 365 employees from several supervisory levels within a public sector organization (the gender breakdown was not provided). Two sets of performance ratings (gathered from both immediate supervisors and next-higher-level supervisors) served as the criterion measures, with participants' job-performance assessed along six dimensions: Written Communication, Leadership, Interpersonal Relations, Planning and Organizing, Analyzing Problems and Making Sound Decisions, and Oral Communication. Although several criterion composites were constructed, only the results based on the mean of the immediate supervisors' ratings across the sample of 365 employees will be reported here, in order to more closely match the criterion measures used in in-basket studies described in Section 2 of the Literature Review. An inter-rater reliability coefficient of .56 was reported for this criterion composite, based on a subsample of 194 subjects. Correlating the in-basket total score with the criterion composite yielded a corrected validity coefficient of .41 \( (p < .01). \) (The uncorrected validity coefficient was .31, reported to be significant at the .0001 level).

A factor analysis of GMIB scores was also performed, resulting in four interpretable factors: Leadership Style and Practices, Handling Priorities/Sensitive Issues, Managing Conflict/Interpersonal Insight, and Organizational Practices/Management Control. The factor scores across the four dimensions were summed and correlated with the criterion composite, yielding a validity coefficient of .41 \( (p < .01). \)

The GMIB has attempted to use balance the realism of a medium-fidelity design with sound psychometric properties and practical scoring methods. These efforts seem to have been successful, as evidenced by several large-scale contracts using the GMIB for administrative ability assessment and a national database of results from 4500 GMIB participants (R. Joines, personal communication, November, 1991).
ETS Consolidated Fund In-Basket Exercise. This widely-used, objectively scored exercise uses the scenario of a fictitious volunteer community fund in which the participant assumes the role of the Fund's paid director who, in coordinating fund-raising activities, must deal with a small staff and Board of Directors made up of prominent citizens. This high-fidelity exercise uses scoring procedures which follow ETS principles that are applied, through analysis of participants' open-ended written responses, to scoring dimensions derived from factor analyses of a much larger set of in-basket categories (Crooks, 1968). The in-basket performance dimensions measured by the Consolidated Fund In-Basket are: Taking Action Toward Solving Problems, Exercising Supervision and Control, Problem Analyzing and Relating, Communicating in Person, Delegating, Scheduling Systematically, Productivity (Amount of Work Accomplished), Quality of Actions Taken, and Scorer's Rating of Overall Performance. The Quality of Actions dimension is the only qualitative or evaluative dimension of the nine because its scoring key is based on the panel approach of industry experts who evaluate all possible courses of action and assign scoring weights for the appropriateness of each action. The Scorer's Rating is impressionistic (raters use a 1 to 5 scale to indicate how well they believe the participant would perform on the job), whereas the remaining seven dimensions are stylistic and are measured by the frequency of occurrence (i.e., quantitative assessment) of actions which involve those stylistic dimensions. The ETS scoring scheme also uses the "Reasons for Action" form, which asks participants, upon completion of the in-basket, to record what they did, and why they chose those actions for each item. As described in the school administration study (Hemphill et. al., 1962), information from the Reasons for Action form is applied to several different dimensions, although the exact keying onto dimensions and weighing of responses is not known.
Very little published empirical information on the Consolidated Fund In-Basket Exercise is available. Even ETS itself has no prepared technical reports on the development and systematic criterion-validation of this popular instrument (K. Abbey, personal communication, June 3, 1992). However, the results from two studies have been published, although they are relatively small-scale applications of the exercise. The first study involved a comparison of in-basket performance across four managerial groups from several companies: 101 high-potential middle managers, 96 lower-middle managers, 34 upper-middle managers, and 175 MBA students (Crooks & Slivinski, 1972). A multiple-group discriminant analysis of scores revealed that while the overall profiles of the four groups were significantly different statistically, subsequent analyses showed considerable overlapping of subgroups within each of the four groups. A series of two-group discriminant analyses were then performed, comparing in-basket performance of the high-potential group with each of the three other groups. As expected, the results indicated that the administrative performance of the high-potential group was most like that of the upper-middle managerial group whereas it was least like performance by the MBA students.

The second study of the Consolidated Fund In-Basket Exercise (French version) examined inter-rater reliability in a sample of 38 French-speaking middle-level managers. Two ETS-trained scorers independently scored each exercise along the nine dimensions previously listed. The correlations across these dimensions ranged from .54 (the impressionistic Scorer's Rating of Overall Performance) to .93 (the more objective Productivity dimension). Interestingly, despite a detailed scoring manual which limits the influence of personal judgement, the important Quality of Actions dimension yielded the second lowest reliability coefficient ($r = .75$).

An example of the combination scoring method. An additional key corporate player in the in-basket field is Development Dimensions International (DDI). Whereas
the MCIME and the GMIB use an objective or psychometric scoring method, the scoring approach used by DDI in both its off-the-shelf and custom-designed in-basket exercises is the combination method, with both an objective and a subjective (interview) component. A large component of DDI's work is the development of tailor-made in-basket exercises for assessment centre applications, many of which are computer-scored. However, neither descriptions of the development nor empirical results of the efficacy of these or other off-the-shelf instruments has been published (A. Smith, personal communication, June 4, 1992).

A description of DDI's off-the-shelf Sellmore Manufacturing Company Foreman's In-Basket (DDI, 1978) illustrates (albeit briefly) the combination scoring method. This high-fidelity in-basket exercise presents items on varying sizes of company stationery and provides blank letterheads for free-format written responses by the participant. The in-basket performance dimensions measured by the instrument are: Sensitivity, Initiative, Planning and Organizing, Analysis, Judgement, Decisiveness, Delegation, and Management Control. In scoring in-basket performance a manual outlining the protocol for the interview is used. The manual also contains several preliminary and concluding interview questions, such as "What did you think of the in-basket," and "What are the major problems confronting you?" In addition, the scorer/interviewer is directed to examine the written responses for each item and discuss each response with the participant in order to decide how many of the "mandatory dimensions to be evaluated in an item" (listed for each item in the manual) were exemplified in the response. If the action(s) taken for the item show evidence of the mandatory dimension, a +1 score toward that dimension is awarded. On the other hand, if no evidence of action(s) exhibiting the mandatory dimension is shown, a -1 score is given for that dimension. The manual also lists non-mandatory dimensions for each item and, unlike the scoring of mandatory dimensions, lack of action on non-
mandatory dimensions does not imply negative behaviour (absence of action demonstrating that dimension is not awarded a negative evaluation).

After reviewing the written response for an item, the scorer/interviewer questions the participant as to why a particular action was taken. The interview therefore allows information similar to what is gleaned from the ETS Reasons for Action Form to be collected (i.e., what was done, and why) but without using the Form's paper-and-pencil format. A primary advantage of the interview is that it provides a greater richness of performance information not measurable through objective means, such as the evaluation of Oral Communication skills (e.g., eye contact, persuasiveness, clarity of expression).

The final step in scoring in-basket performance using this example of the combination method is the scorer's completion of the Overview of In-Basket and In-Basket Interview. For each dimension, a "Dimension Summary Sheet" contains an item grid of all items across the exercise in order to record summaries of actions that demonstrate that dimension. An overall rating of the effectiveness of the participant's performance is made in the particular metric applied by individual assessment centres. The manual acknowledges that usually a 0 to 5 scale is applied, with 0 indicating that no opportunity existed for the dimension to be shown and 5 indicating that a great deal of the dimension was shown. Lastly, a subjective summary is made of the major strengths and weaknesses of the participant's in-basket performance.

Summary of current in-basket scoring methodology in industry. There is a marked lack of information regarding instrument descriptions, scoring protocols, and empirical evaluations of current in-basket scoring practices in the field. For example, no known published information exists regarding instruments solely using the subjective method of scoring and important detail is often missing in reports of in-basket exercises.
using the objective and combination scoring methods. Moreover, as noted earlier, an accurate assessment of the relative rates of use of the three scoring methods in industry is not available. According to Thornton, "...no good data exists...we simply don't know about in-basket scoring practices in industry" (G. Thornton, personal communication, June 30, 1992). Although little is known about general scoring practices in the field, several specific examples of widely-used in-basket exercises were discussed in order to illustrate (as fully as possible) traditional industry applications (e.g., The Consolidated Fund In-Basket Exercise), as well as newer, less conventional in-baskets which incorporate recent design and scoring innovations (e.g., The MCIME).

Summary of Section 3: A Review of Current In-Basket Scoring Strategies

This third and final section of the Literature Review began by elaborating several limitations of a literature-based review of current in-basket scoring strategies, including the absence of a comprehensive classification system of in-basket scoring methods. Accordingly, the present work has provided a scoring classification system, first outlined in the Introduction and briefly reviewed in this section. Consideration was then given to current industry-based in-basket designs and the scoring practices used to evaluate administrative performance. Frederiksen's observation, made two decades ago, that "the work of the industrial psychologist often does not find its way into the psychological literature" (1972, p. 69) continues to be valid regarding in-basket design, as a whole, and in-basket scoring, in particular. The reporting of in-basket design and scoring must become more common and standardized in order to allow an accurate assessment of the efficacy (psychometrically and practically) of various scoring methods.
PRELIMINARY STUDIES

Study 1

Method

Participants and Setting

The sample consisted of 321 first-level supervisory employees (entry-level management personnel) from different departments of a large Canadian utility company (168 males; 153 females). In the fall of 1989, participants were given a newly-revised version of an existing in-basket exercise as part of a concurrent validation study for a managerial assessment battery containing intellectual, personality, biodata, and supervisory judgement measures. The full assessment battery required 8 hours to complete and was administered in a 1-day session. Participants were given 1 1/2 hours to complete the in-basket exercise, preceded by 15 minutes of scripted instructions read aloud to enhance both the clarity and standardization of exercise instructions.

Materials

The Telephone Supervisor In-Basket Exercise (TSIB). This in-basket exercise consisted of 21 items or problems to be dealt with by the fictitious character Chris Wilson, who was recently promoted to First-Level Supervisor for the Independent Telephone Company of Iowa. The scenario required Chris to handle, on a Sunday afternoon, the work that had accumulated in the in-basket of the new position, 90 minutes prior to leaving with the new boss to attend the yearly Budget Meeting. The in-basket items included an angry letter from a dissatisfied customer, internal memos suggesting a large backlog for residential and industrial telephone service, evidence of budget overruns, and staff conflicts. Considerable effort was made to enhance the
realism of the stimulus materials and response format (i.e., fidelity). Accordingly, company stationery of varying sizes and design, complete with the company logo, were provided. Participants were instructed to take actions as necessary and accomplish as much as they could in the time available. They were told to write out memos and letters, to sign papers, if necessary, and to use paper clips provided to attach their response memos and letters to the relevant items.

Criterion measurement. A combination of three behaviourally-anchored rating scales (BARS; Smith & Kendall, 1963) and three behavioural observation scales (BOS; Latham & Wexley, 1981) developed specifically for this study were used for each dimension. A total of 12 performance appraisal dimensions were measured by the participants' immediate supervisor. These appraisal dimensions were developed by the company from an exhaustive list of behavioural incidents generated by 22 company managers. The specific appraisal measures are described more fully in an article by Hakstian, Woolley, Woolsey and Kryger (1991). They comprised Leadership, Planning/Organizing/Control, Oral Communication, Analysis, Judgement, Decisiveness, Work Ethic, Initiative, Behaviour Flexibility, Sensitivity, Performance Stability, and Written Communications. The supervisors providing criterion ratings were trained in the meaning of the definitions used for the criteria and were given clear instructions for the proper completion of the appraisal forms. Each participant was measured on each of the 12 performance dimensions by means of three BARS and three BOS, for a total of six scales. The average time required to complete a rating on one participant was 1 1/2 hours. Assessment battery and performance criterion scores were collected concurrently.

The first principal component of the 12 performance dimension scores was then calculated and called the Overall Performance Criterion (OPC). Because it is a weighted average of the performance dimensions with the added advantage of maximal
internal consistency, this criterion measure is seen as the best measure of overall management performance.

Procedure

Development of the TSIB. An existing in-basket exercise described in Hakstian et al. (1986) was modified in order to improve its reliability and validity. After a review of the literature, several steps were taken to change the original instrument, including modifications to item design and response format of the in-basket, as noted earlier.

Firstly, it was apparent that the ratio of number of items per hour in the original exercise was much higher than that typically found in other in-baskets. Accordingly, the time allotted for the exercise was extended from 1 to 1 1/2 hours to fit the typical ratio of approximately 15 items per hour. It was hypothesized that such a change would allow greater breadth and depth of responses, and so help to yield richer information along a greater number of performance dimensions.

A further focus for change was to improve the realism of the response format. As we have seen in the Literature Review, realism is necessary to provide motivation for the participants to take the exercise seriously and become "ego-involved" (Lopez, 1966). Thus, in order to allow female participants to adopt the role and fit in more easily, the male persona portrayed in the original in-basket was changed to a gender-neutral persona. Similarly, the structure of the response format was altered to provide more realism; instead of using an "action booklet" wherein responses were recorded in a small bound notebook, company stationery of varying sizes and design were introduced. Participants were instructed to use paper clips to attach their response memos and letters to the relevant items. Moreover, the original in-basket required participants to determine the exact order in which they would complete all items, as a
means of assessing the ability to recognize the priority of items. It was felt that such a determination was somewhat artificial and may over-emphasize a cognitive, rather than action-oriented, approach. Again, given greater freedom in response format, allowing more creativity and uniqueness of responses, and given more time in which to show administrative acumen, it was hypothesized that more accurate and valid data would emerge.

Such positive changes, however, would not be without consequence. In essence, the main impact of such modifications was to increase both the complexity of, and time required for, scoring. In addition, a more complex scoring system could result in decreased reliability with an attendant reduction in validity. Thus, in conjunction with these changes to item design and response format, an additional approach aimed at modifying the scoring system was designed and incorporated in order to minimize possible deleterious psychometric effects.

Identification of the TSIB dimensions. This additional approach involved a series of changes designed to make the scoring system as objective, reliable and valid as possible. In the summer of 1989, a large Canadian utility company using the in-basket described in Hakstian et al (1986) in its management selection program agreed to make available 60 randomly chosen in-baskets completed by managerial candidates. The author compiled a list of the responses for all 21 in-basket items (and across all 60 in-baskets) in order to identify the typical courses of action taken and thus derive a list of action elements for each item. (It will be recalled from the Literature Review that action elements are the smallest, distinguishable units of action which result from breaking down in-basket responses.)

Later that summer, the author met with Lois Crooks and Peggy Mahoney of Educational Testing Service (ETS), whose research department sponsored Frederiksen's
(1957, 1962) pioneering work, for instruction in the development of an objective in-basket scoring system. This consultation led to the identification of the dimensions to be used in the newly-revised in-basket, which were operationalized as follows:

1. **Planning and Organizing Work:** This dimension included organizing work according to related content or priority, and scheduling systematically for definite times.

2. **Interpersonal Relations:** This dimension reflected the degree of sensitivity, interpersonal awareness and skills shown.

3. **Leadership in a Supervisory Role:** This dimension indicated the extent to which the participant assumed leadership, i.e. took the lead in solving problems and in carrying out policies set by superiors, changing procedures, deciding critically whether to comply with the suggestions or proposals of others.

4. **Managing Personnel:** This dimension involved supervision and controlling the activities of staff, asking for information, resolving staff conflicts, and giving directions and suggestions to personnel.

5. **Analysis and Synthesis in Decision-Making:** This dimension reflected the extent to which the participant analyzed problems through means such as taking into account available information and considering policy or other aspects of problems.

6. **Productivity:** This dimension was measured by the number of actions taken on all items and the number of items attempted.
7. **Quality of Judgement:** This dimension reflected the appropriateness of actions taken, based on pooled opinions of experienced persons who had previously considered the range of the possible courses of action.

**Development of TSIB dimension scoring keys.** In the late summer of 1989, the scoring key for Quality of Judgement was generated using the standard ETS practice of determining the plurality panel ratings of appropriateness for each action element.

However, before this method is described, an important development to improve the validity of the crucial Quality of Judgement dimension should be acknowledged. As described in the Literature Review, a panel of industry experts is typically constructed in order to determine the Quality of Judgement scoring key by assigning ratings of appropriateness (or judgement) for each action element. The development introduced here involved increasing the size of the panel based on the hypothesis that a larger, company-specific set of ratings would increase the stability and validity of the judgement ratings. Whereas the original in-basket used a three-member panel (two human resource employees and a senior researcher) to assign appropriateness ratings to action elements, the new version of the in-basket used an 11-member panel of mid-to-senior level managers from the target company (later to referred to as Panel 1). Despite efforts to construct a gender-balanced panel, only 1 of the 11 panel members was female.

The process of determining the Quality of Judgement scoring key involved a 1-day workshop with the entire 11-member panel. In order to familiarize the panel with the in-basket exercise, the first part of the workshop required each panel member to actually complete the in-basket exercise following the standardized administration procedure. Next, each member was given a prepared list of action elements for each item (derived from the 60 in-basket exercises referred to in the preceding Identification
of the TSIB dimensions section). Each panel member was then asked to consider each element independently and assign a value of -1, 0, or +1 to reflect whether that element would be an unfavorable, neutral, or favorable step toward solving the problem presented in the item. After the workshop, the author tallied the scoring assignments for each action element across all panel members and determined final scoring weights by plurality of judgements. For example, if six panel members assigned a +1, whereas five members assigned a -1, that action element would be keyed positively. If five members assigned a -1 to an action element, while four rated it +1, and the remaining members gave it a 0, the action element would be keyed negatively.

The remaining five dimensions, from Planning and Organizing Work to Analysis and Synthesis in Decision-Making, are seen as stylistic dimensions because they are descriptive measures of the manner in which a participant tends to respond. The process of determining the scoring keys for these dimensions was based on the ETS procedure of considering each action element and logically assigning the dimensions to elements. Specifically, the author and two colleagues made independent assignments based on the perceived relevance of each dimension to each action element. The author then made the final decisions regarding the assignment of the dimension(s) to action elements. For the majority of action elements, multiple dimension assignments were made. Figure 3 presents a sample item from the in-basket exercise used in both the preliminary and present studies. Figure 4 displays an excerpt from the Scoring Manual for the same item, listing the action elements and outlining the Quality of Judgement key and remaining stylistic dimensions coded for each action element.

**General scoring protocol.** Scoring of the 321 in-basket exercises was conducted by three employees from the target company and two external consultants who underwent a 5-day training process directed by the author. Scorers used a detailed
Scoring Manual prepared by the author (110 pages) which provided a description of the in-basket exercise, general procedures and guidelines for scoring, and a list of the action elements and scoring keys for each item in the in-basket exercise (as noted, an illustration of the action elements and scoring key for one item has been provided in Figure 4). In total, the Manual listed approximately 600 action elements (with assigned dimensions for scoring) for the 21 items, with an average of 29 elements per item.
MEMORANDUM

To: Reed
From: Hal
Subject: Coffee Breaks
Date: 9/1

It has been brought to my attention that for the last three weeks, four of your trucks have been parked at Raymond's Coffee Shop on Parks Ave. for about 45 minutes, instead of 15, each afternoon. We are already behind in due dates, so let's get this corrected!

Figure 3. Sample Item from the Telephone Supervisor In-Basket Exercise.
1. Item 4: Memo from Hal regarding trucks parked by Coffee Shop

<table>
<thead>
<tr>
<th>Quality of Judgement</th>
<th>Other Dimensions</th>
<th>Action Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ 3,4,5</td>
<td></td>
<td>1. Plan to find out who is involved</td>
</tr>
<tr>
<td>+ 2,3,4,5</td>
<td></td>
<td>2. Give reprimands to those involved</td>
</tr>
<tr>
<td>0 2,3,4</td>
<td></td>
<td>3. Call to attention of Service Centre informing someone of break problem</td>
</tr>
<tr>
<td>+ 3,4,5</td>
<td></td>
<td>4. Plan to drive by Coffee Shop</td>
</tr>
<tr>
<td>+ 1</td>
<td></td>
<td>5. Defer until return (no other action)</td>
</tr>
<tr>
<td>+ 1,5</td>
<td></td>
<td>6. Note urgency (do immediately, ASAP)</td>
</tr>
<tr>
<td>+ 2,3,4</td>
<td></td>
<td>7. Plan to motivate employees to work harder and more efficiently by talking to them in friendly, positive way - must be directly stated</td>
</tr>
<tr>
<td>0 2,3</td>
<td></td>
<td>8. Memo to Hal that Wilson is looking into it (no mention of follow-up or further contact)</td>
</tr>
<tr>
<td>+ 1,2,3</td>
<td></td>
<td>9. Plan to discuss with Hal in car on way to Budget Meeting/upon return</td>
</tr>
<tr>
<td>+ 2,3</td>
<td></td>
<td>10. Plan to discuss with Hal (time not specified)</td>
</tr>
<tr>
<td>0 2,3</td>
<td></td>
<td>11. Memo to Hal about corrective action (not specified)</td>
</tr>
<tr>
<td>+ 2,3,4,5</td>
<td></td>
<td>12. Memo to truckers to clarify 15 minute, not 45 minute breaks</td>
</tr>
<tr>
<td>+ 1,5</td>
<td></td>
<td>13. Relate to Item 20 (letter from customer about lack of service)</td>
</tr>
<tr>
<td>- 2,3,4,5</td>
<td></td>
<td>14. Memo to Wayne Freed threatening disciplinary action if further complaints of extended breaks are received</td>
</tr>
<tr>
<td>+ 1,5</td>
<td></td>
<td>15. Relate to Item 10 (Proficiency/Productivity memo from Hal)</td>
</tr>
<tr>
<td>- 2,3,4</td>
<td></td>
<td>16. Add to Item 10 memo that coffee breaks are not to be abused</td>
</tr>
<tr>
<td>+ 1,5</td>
<td></td>
<td>17. Relate to Item 3 (Personnel Record)</td>
</tr>
</tbody>
</table>

Figure 4. Excerpt from the Scoring Manual for Sample Item (shown in Figure 3)
The primary task of the scorer was to match, as closely as possible, the participant's written response to the most appropriate action element listed in the Scoring Manual for the item. In keeping with standard ETS scoring procedures, provision for "Unusual" or imaginative responses was made (those not provided in the list of typical action elements). In such cases, the scorer had to first decide whether the response was similar enough to one listed in the Manual. If not (i.e., the response was truly unusual), the scorer then had to decide whether the response was scorable and construct a new action element for that response. These responses were then credited or assigned under Productivity and any other dimensions deemed relevant by the scorer, but no Quality of Judgement value was assigned because no panel had evaluated the appropriateness of these novel actions.

Results

Reliability

Criterion measurement. A serious consequence of unreliability of criterion measurement is the attenuation of validity coefficients. Much effort was devoted to minimizing the possibility of rater error and bias in the development and application of the criterion measures. A generalizability analysis was carried out on the scales used to measure the performance dimensions. The average generalizability coefficient of the twelve performance dimensions measured by the six scales was .83 (with each assessed in a two-facet generalizability design).

TSIB dimensions. The three trained company scorers independently scored 39 in-baskets (19 males; 20 females) for the seven dimensions outlined previously. Table 2 shows the inter-rater reliabilities for the dimensions, with coefficients ranging from
Table 2

Inter-Rater Reliability Estimates for TSIB Dimensions

<table>
<thead>
<tr>
<th>In-Basket Dimension</th>
<th>Inter-Rater Reliability&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Planning and Organizing Work</td>
<td>.95</td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td>.82</td>
</tr>
<tr>
<td>3. Leadership in a Supervisory Role</td>
<td>.82</td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td>.87</td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td>.92</td>
</tr>
<tr>
<td>6. Productivity</td>
<td>.94</td>
</tr>
<tr>
<td>7. Quality of Judgement</td>
<td>.91</td>
</tr>
</tbody>
</table>

Note. These results are based on a reliability subsample of 19 males and 20 females.

<sup>a</sup> This is the inter-rater reliability estimate, by analysis of variance, for a single rater.
.82 to .95. These high coefficients suggest the raters were in strong agreement in scoring the action elements measuring the seven in-basket exercise dimensions. The internal consistency reliability of the seven in-basket dimensions was also calculated, separately by gender, yielding values of .87 (males) and .85 (females). The seven dimensions, therefore, are relatively highly correlated.

Validity

Table 3 presents the concurrent validity coefficients between the Overall Performance Criterion, OPC, and the in-basket dimension scores. Validity results presented in Table 3 are somewhat mixed; two of seven dimensions showed low, but statistically significant, relationships with overall performance for both genders, with Quality of Judgement for the males also contributing useful information.

Discussion

In summary, the results obtained in Study 1 revealed that the scoring system for the seven dimensions was highly reliable, and that certain TSIB dimension scores were significantly related to job performance ratings. Validity results, however, were very modest, especially considering previous findings reported in the literature (Brass & Oldham, 1976; Kesselman et al., 1982). It should be noted that, on average, one in-basket required between 1 1/4 and 1 1/2 hours to score. In short, the exercise required as long to score as it did to administer.

The modest validity findings of Study 1 (particularly for the Quality of Judgement dimension), as well as company concerns of lengthy scoring time, supported the continued development of the TSIB. Future research to assess the impact of modifications to the scoring keys appeared warranted; in this way, it was hypothesized that psychometric and practical improvements could be realized. Specifically, a
<table>
<thead>
<tr>
<th>In-Basket Dimension</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Planning and Organizing Work</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td>.21</td>
<td>.15</td>
</tr>
<tr>
<td>3. Leadership in a Supervisory Role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td>.21</td>
<td>.21</td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Productivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Quality of Judgement</td>
<td></td>
<td>.20</td>
</tr>
</tbody>
</table>

**Note.** Reported $r$'s have been corrected for effects of range restriction and criterion unreliability. All are significant at the .05 level (one-tailed tests). All non-significant correlations were omitted.
promising area for further investigation would be to examine the effect of several scoring keys, each constructed differently, on the validity of Quality of Judgement scores. Because this dimension is regarded as the most critical in terms of understanding managerial potential, it would be the logical choice as the focus for scoring key development.

An additional avenue for future research would be to assess the impact that scoring a *subset* of items, rather than the complete set of 21 items, would have on the validity of Quality of Judgement scores. In addition to reducing scoring time, it seemed plausible that eliminating those items which showed negative or low validity coefficients with the criterion, as well as deleting those which showed inconsistent results across genders, would result in an increase in validity coefficients. It should be noted that although fewer items would be scored, participants would still be administered the full exercise, and they would have no knowledge of which items were to be selected for scoring.

**Study 2**

**Method**

**Participants and Setting**

Three new scoring keys to measure the Quality of Judgement dimension were constructed and applied to the same dataset from Study 1. Moreover, three different subsets of items consisting of 8, 10 and 12 items were selected and the results from the 321 subjects were rescored on these three subsets of items.
Procedure

Modification of the Quality of Judgement scoring key. In an effort to increase the validity of the Quality of Judgement dimension, a re-analysis of all Unusual responses across the 21 items in Study 1 was undertaken by the author. It should be recalled that Unusual action elements carry no Quality of Judgement rating, and possible inconsistencies (hence error) would be introduced if a scorer is to decide whether an action is or is not scorable (i.e., listed in the action element list provided in the Scoring Manual). Thus, to glean as much scorable information as possible about participants' responses and to reduce possible error from scorers' judgements, unusual responses from the 321 subjects were studied. When three or more participants endorsed the same unusual action, it was then included on the new list of action elements for that item. Logical keying of the stylistic dimensions was also made for each "new" action element. What remained was to determine and assign Quality of Judgement ratings for these new elements.

A new, expanded panel. A new, second panel was struck at a different company whose work was very similar to that of the first company. The same 1-day workshop format described in Preliminary Study 1 was followed here. To increase the accuracy and soundness of the appropriateness ratings even further, 21 judges were used in total for this second panel, compared to 11 for the first panel used in Study 1. Because of personnel requirements, however, not all of the selected panel members could attend the 1-day panel workshop at the same time. As a result, the 21-member panel was divided into two sub-groups of a 9- and a 12-member panel. Ratings collected from these two sub-panels from the two workshops were then combined to yield the overall 21-member evaluations. The first 11-member panel and the second 21-member panel are referred to as Panel 1 and Panel 2, respectively, reflecting the origins of the key from Company 1 or Company 2. Size is not the most useful
classification basis for the panels because, as will be shown, not all action elements from Company 2 were rated by 21 members (as noted, some were based on the 12-member sub-panel).

The establishment of Panel 2 (from Company 2) allowed the newly-added action elements from the Unusual action re-analysis to be rated. However, the re-analysis and recoding of Unusual responses into action elements described earlier was completed after the first nine-member panel had met. The second sub-panel of 12 members was then provided with the new, expanded list of action elements to make their ratings of appropriateness of actions. Thus, although most of the elements were rated based on a 21-member panel, some of the newly-added action elements were based on judgements from the 12-member sub-panel.

**Derivation of the new scoring keys for Quality of Judgement.** Panel 2 ratings also provided the opportunity for several new Quality of Judgement keys to be generated. For the first key, a variation of the original panel method of deriving overall panel ratings was introduced. Labelled the Panel 2 (Conservative) key, the main distinction from the original method used to derive the Panel 1 key was the application of more conservative guidelines which required a higher proportion of agreement before a positive or negative value would be assigned. Specifically, instead of a simple majority, 75% panel agreement was needed before a maximum value of +2 or -2 would be given. Between 50% and 75% agreement resulted in either a +1 or -1 (as long as 25% or fewer marked the opposite sign), and zero would be assigned for all other outcomes. Thus, the metric was also moderately expanded from a three to five point scale to allow finer gradations of agreement. This same method of using a 50% and 75% cutoff was used for those elements based on judgements from the 12-member sub-panel. For clarification, the convention of indicating the method of combining panel judgements will now be followed. Specifically, the Panel 1 key will be referred
to as Panel 1 (Liberal) key, and the Panel 2 key will be referred to as the Panel 2 (Conservative) key.

The second key, called the Empirical key, followed Meyer's (1970) correlational analyses to derive scoring weights in that the endorsement of each action element was correlated with the OPC by means of a point-biserial correlation coefficient. Action elements were treated as dichotomous predictor variables; where an action element was endorsed, the candidate received a score of 1, whereas his/her action element score for unendorsed elements was zero. The criterion measure, OPC, was a continuous variable. If those rated higher on the overall performance scale tended to endorse an element like "meeting with a subordinate" while lower performers tended not to take such an action, we would see a positive correlation, and that element would be keyed positively. If poorer performers chose an action element while the better performers did not, we would see a negative correlation and that element would be keyed negatively. If no clear pattern between management performance level and endorsement of action elements was present, we would see negligible correlations, and those action elements would receive a neutral value.

More specifically, the Empirical key decision rules for assigning quality of judgement values considered both the size and direction of the observed correlations, as well as gender (the correlations from both genders had to agree in sign in order for a positive or negative value to be assigned). In effect, each gender acted as a cross-validity check on the other. The specific guidelines used in the Empirical key to assign values to all action elements were as follows:

1. \[ +2 \text{ or } -2: \] separate-gender correlations agreed in sign and the product of \( p \) values was less than or equal to .05;
2. +1 or -1:

(a) separate-gender correlations agreed in sign and the product of $p$ values was greater than .05 and less than or equal to .24;

OR

(b) separate-gender correlations differed in sign and:

i) the larger correlation had a $p$ value less than or equal to .10 and the smaller correlation had a $p$ value of greater than .30

OR

ii) the larger correlation had a $p$ value less than or equal to .01 and the smaller correlation had a $p$ value larger than .20

For (b), the sign keyed was the same as that of the larger correlation

3. 0 for all other outcomes

It should also be noted that zero correlations were counted as agreeing in sign with the other, non-zero, correlation and a $p$ value of 1 was assigned to all zero correlations. Lastly, correlations based on endorsements across the sample of less than three participants were omitted.
The third key, called the Merged key, was a complex combination of the Panel 2 (Conservative) and Empirical keys. It included consideration of both the correlational results (size, significance level, and directions for each gender) as well as overall Panel 2 (Conservative) ratings. Specifically:

1. +2 or -2:
   (a) sign in panel was matched in both genders, and the product of the \( p \) values was less than or equal to .12;
   (b) in the case of the panel having assigned a zero, both gender results were in the same direction with the product of \( p \) values less than or equal to .06;

2. +1 or -1:
   (a) sign in panel was matched in both genders and the product of the \( p \) values was greater than .12 and less than or equal to .24;
   (b) sign in the panel agreed for one gender result, with a \( p \) value less than or equal to .12 (this sign is keyed), and the other gender result (in the other direction) had a \( p \) value of greater than .20;
   (c) when panel = 0, both gender results were in the same direction, with the product of the \( p \) values larger than .06 and less than or equal to .15;
   (d) sign in panel was opposite to both gender results (which were in the same direction), and the product of the \( p \) values was less than or equal to .05 (keyed in direction of the empirical results);
(e) when the panel = 0, gender results were in opposite directions, with one gender result (one keyed) with a $p$ value of less than .06 and the other (opposite) result with a $p$ value of greater than .25.

3. 0 for all other outcomes

Here, zero correlations were counted as agreeing in sign with either the panel or the other gender result. If a need arose to choose between the panel and the gender result, the correlation was seen as agreeing with the gender result. A $p$ value of 1 was given to all zero correlations for computations. As with the Empirical key, correlations based on endorsements across the sample of less than three participants were omitted.

**Derivation of the reduced-item sets.** To select the best items for the 8-, 10-, and 12-item subsets, item analyses were conducted using the item-reliability index and the item-validity index. The item-reliability index is the product of the standard deviation of the item and the correlation between the item score and the total test score. The item-validity index is the product of the standard deviation of the item and the correlation between the item score and the criterion score. Reliability and validity indices were calculated for each of the 21 items in the TSIB using the data from Study 1 for each gender and the Quality of Judgement scores as measured by the Panel 1 (Liberal) key. Final selections of the components of the three subsets of items were made on the basis of consistency in results across genders and on the basis of the items' relative contributions to validity, and, to a lesser degree, reliability.

**Results**

**Reliability**

In Table 4, internal consistency reliability results appear for the Quality of Judgement scores obtained from the three subsets of items as scored by the Merged
key. Although the internal consistency reliabilities reported in Table 4 show moderately low reliabilities, the split-half reliability estimates are more promising than those obtained by the Panel 1 Liberal method applied to the 21-item set (these are reported under the Note in Table 4).

Validity

Table 5 contains the results from correlating the Quality of Judgement scores calculated using several scoring keys (Panels 1--Liberal and 2--Conservative, Empirical, and Merged) and three combinations of items with the Overall Performance Criterion score. It should be clearly understood that Panel 1 refers to that scoring key derived from the first panel of 11 members from the Company 1 whereas Panel 2 refers to ratings based on Company 2's panel of 21 members.

Table 5 shows, not surprisingly, that validity coefficients are strongest for the Empirical key. The contrast in magnitude of coefficients between both Panel keys and the Empirical key is marked. It should also be noted that incorporating Panel 2 (Conservative) judgements with the Empirical results to yield the Merged key did not result in significantly lower validities. It should be made clear that the item composition of the item sets measured (8-, 10- and 12-items) is exactly the same for each scoring key applied.

Intercorrelations

Table 6 contains the results of the intercorrelations among the seven dimension scores based on the 21-item set whereas Table 7 provides the intercorrelations based on the 8-item subset, using the previously described logical dimension keying system for scoring these stylistic dimensions, and the Panel 1 (Liberal) key to score the Quality of Judgement dimension. The fairly high level of intercorrelation among the seven
Table 4

Alpha and Stepped-up Split-Half Reliability Estimates for Quality of Judgement Scores Obtained from Three Item Sets as Scored by the Merged Key

<table>
<thead>
<tr>
<th>Item Sets</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td>.45</td>
<td>.40</td>
</tr>
<tr>
<td>10 item set</td>
<td>.53</td>
<td>.54</td>
</tr>
<tr>
<td>12 item set</td>
<td>.56</td>
<td>.57</td>
</tr>
<tr>
<td><strong>Split Half</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td>.52</td>
<td>.41</td>
</tr>
<tr>
<td>10 item set</td>
<td>.58</td>
<td>.58</td>
</tr>
<tr>
<td>12 item set</td>
<td>.65</td>
<td>.62</td>
</tr>
</tbody>
</table>

Note. For comparison, the alpha reliability coefficients for the 21-item set, scored by the Panel 1 (Liberal) key, were .54 (males) and .60 (females). The stepped-up split half reliability coefficients for the 21-item set, scored by the Panel 1 (Liberal) key, were .28 (males) and .48 (females).
Table 5

**Bivariate Correlations between Overall Performance Criterion and Quality of Judgement Scores Obtained from Four Scoring Keys and from Three Item Sets**

<table>
<thead>
<tr>
<th>Item Sets</th>
<th>Panel 1</th>
<th>Panel 2</th>
<th>Empirical</th>
<th>Merged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Liberal)</td>
<td>(Conservative)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Males</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td>.37*</td>
<td>.29*</td>
<td>.61**</td>
<td>.60**</td>
</tr>
<tr>
<td>10 item set</td>
<td>--</td>
<td>.23*</td>
<td>.64**</td>
<td>.62**</td>
</tr>
<tr>
<td>12 item set</td>
<td>--</td>
<td>.23*</td>
<td>.63**</td>
<td>.60**</td>
</tr>
<tr>
<td><strong>Females</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td>.29*</td>
<td>.31*</td>
<td>.63**</td>
<td>.59**</td>
</tr>
<tr>
<td>10 item set</td>
<td>--</td>
<td>.28*</td>
<td>.62**</td>
<td>.62**</td>
</tr>
<tr>
<td>12 item set</td>
<td>--</td>
<td>.25*</td>
<td>.63**</td>
<td>.62**</td>
</tr>
</tbody>
</table>

**Note.** Reported $r$'s have been corrected for effects of range restriction and criterion unreliability. Correlations for the 10- and 12-item sets, scored by the Panel 1 (Liberal) key, are not available.

$p < .01$ (one-tailed). **$p < .001$ (one-tailed).
Table 6

Intercorrelations of Dimension Scores based on the 21-Item Set: Stylistic Dimensions Scored by Logical Keying System and Quality of Judgement Scored by Panel 1 (Liberal) key

<table>
<thead>
<tr>
<th>Dimension</th>
<th>P &amp; O</th>
<th>IR</th>
<th>LSR</th>
<th>MP</th>
<th>A &amp; S</th>
<th>P</th>
<th>QJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning and Organizing Work</td>
<td>.39</td>
<td>.30</td>
<td>.24</td>
<td>.77</td>
<td>.87</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.79</td>
<td>.79</td>
<td>.11</td>
<td>.54</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.65</td>
<td>.19</td>
<td>.50</td>
<td>.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td></td>
<td>.02</td>
<td>.44</td>
<td>.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.76</td>
<td>.52</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.67</td>
</tr>
<tr>
<td>Quality of Judgement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension</th>
<th>P &amp; O</th>
<th>IR</th>
<th>LSR</th>
<th>MP</th>
<th>A &amp; S</th>
<th>P</th>
<th>QJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning &amp; Organizing Work</td>
<td>.28</td>
<td>.19</td>
<td>.25</td>
<td>.74</td>
<td>.84</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.78</td>
<td>.79</td>
<td>-.07</td>
<td>.39</td>
<td>.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.62</td>
<td>-.03</td>
<td>.32</td>
<td>.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>-.05</td>
<td>.39</td>
<td>.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis &amp; Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.78</td>
<td>.48</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.69</td>
</tr>
<tr>
<td>Quality of Judgement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7

Intercorrelations of Dimension Scores based on the 8-Item Set: Stylistic Dimensions Scored by Logical Keying System and Quality of Judgement Scored by Panel 1 (Liberal) key

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P &amp; O</td>
<td>IR</td>
</tr>
<tr>
<td>Planning &amp; Organizing Work</td>
<td>.36</td>
<td>.32</td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.76</td>
<td>.77</td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.56</td>
<td>.34</td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>.13</td>
<td>.46</td>
</tr>
<tr>
<td>Analysis &amp; Synthesis in Decision-Making</td>
<td>.84</td>
<td>.38</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of Judgement</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Planning & Organizing Work              | .21   | .14 | .22 | .64 | .71   | .45 |
| Interpersonal Relations                 | .77   | .69 | .04 | .51 | .50   |
| Leadership in a Supervisory Role        | .54   | .11 | .43 | .49 |
| Managing Personnel                      | .17   | .48 | .54 |
| Analysis & Synthesis in Decision-Making | .79   | .53 |
| Productivity                            |       |     |     |     | .67   |
| Quality of Judgement                    |       |     |     |     |       |
dimensions as measured by both the 21- and 8-item sets suggests a lack of meaningful
discriminant validity of several aspects of administrative performance.

Construct validity

Lastly, the conceptual nature of in-basket performance was examined by correlating Quality of Judgement scores as derived by the 10-item set (to choose a "moderately-sized" subset) as scored by Merged key (Key 3) with selected intellectual and personality variables. (A list of the cognitive ability and personality tests used in this analysis is provided in Appendix A.)

It is not appropriate to summarize all findings here, except to briefly describe the strongest relationships between in-basket performance (Quality of Judgement) and the set of external variables examined. All following reported correlation coefficients were tested using one-tailed significance tests. Bivariate correlations with intellectual measures revealed that reading comprehension was related to higher in-basket performance for both men ($r = .20, p < .005$) and women ($r = .26, p < .001$). A stronger pattern emerged for women across additional communication skills like writing ability ($r = .19, p < .01$) and vocabulary ($r = .20, p < .01$). Both genders showed a significant relationship between Quality of Judgement and the ability to think in cognitively flexible ways (males: $r = .20, p < .005$; females: $r = .34, p < .000$).

Correlations with personality measures suggest that, across genders, characteristics such as assertiveness and dominance are moderately related to in-basket performance. Specifically, correlations of .17 (males, $p < .05$) and .20 (females, $p < .01$) were observed when Quality of Judgement was related to dominance. When correlated with assertiveness (16PF Factor E), coefficients of .15 (women, $p < .05$) and .19 (males, $p < .01$) were seen. Gender-specific results suggested that women who were unconventional ($r = .25; p < .001$) and independent ($r = .24; p < .001$)
also seemed to show better administrative judgement. A different pattern emerged for men; those who tended to score higher on the Quality of Judgement dimension were also enthusiastic (16PF Factor F; \( r = .18; p < .01 \)) and extroverted (16PF Second-Order Factor QI; \( r = .17; p < .05 \)).

Discussion

The goals of Study 2 were to further the research begun in Study 1 by (a) examining the effect that three scoring keys, each constructed differently, would have on the validity of the Quality of Judgement scores, and (b) constructing three different subsets of items and assessing the impact of scoring each on reliability and validity.

The three keys were constructed on the basis of three frames of reference:

1. the logical or rational Panel 2 (Conservative) key,
2. the Empirical key, and
3. the Merged key, a combination of the Panel 2 and Empirical keys.

In the derivation of the reduced item sets, the items were selected more for their contribution to validity, rather than reliability. One of the chief concerns with reducing the set of scored items in this way was the possible detrimental effect this would have on the reliability of the instrument by reducing the variability and the overall stability of the scores. It was decided to methodically examine this concern by constructing the three different subsets of items: an 8-item, 10-item, and 12-item scored set of items. Thus, the relative impact of a reduced item set on both reliability and validity could be assessed.
In sum, then, the main question asked in Study 2 was what effect would the three ways of scoring Quality of Judgement applied to three subsets of items scored have on the psychometric properties of the exercise.

Overall, we found confirmation of past research findings that the in-basket is a reliable and valid way of assessing administrative ability, even when the set of items actually scored is reduced substantially and new scoring keys are applied. Scoring time was decreased from 1 1/4 - 1 1/2 hours to 23 minutes, on average, for the 8-item key. Another practical benefit to the modifications examined in Study 2 was the reduction in time required to train scorers, who would be responsible for becoming skilled in scoring less than half of the original items. The time required to train scorers was slightly more than half of what had been required when all items were scored (i.e., three days for the 8-item subset versus five days for the full set of 21 items). Accordingly, it was found that important practical advantages of reducing training and scoring time could be realized without attendant reductions in the reliability and validity of the instrument.

As noted earlier, the internal consistency reliabilities presented in Table 6 showed moderately low reliability coefficients for the Quality of Judgement dimension as based on the three subsets of items as scored by the Merged key, although the split-half reliability estimates were more promising than those obtained by the Panel 1 (Liberal) key applied to the full 21-item set. However, if we were attempting to measure a unitary, factorially-pure trait, the items would have to be seen as lacking the requisite internal consistency. It is possible, however, that the Quality of Judgement dimension is less homogeneous than originally thought; perhaps these items are measuring somewhat independent aspects of administrative behaviour that make up the overall Quality of Judgement dimension. It should also be recalled that, in the derivation of the reduced-item subsets, the items were selected more for their
contribution to validity, rather than reliability, so maximal internal consistency of items was not expected. Across all dimensions (particularly the stylistic dimensions), Tables 6 and 7 showed evidence of a fairly high level of intercorrelation for both the full and a reduced item set, suggesting a lack of meaningful discriminant validity of several aspects of administrative performance.

We have seen that concerns of reduction in validity with the scoring key changes and reduced-item scoring approach examined in this study were not realized. In fact, the Quality of Judgement dimension results shown in Table 5 suggested marked improvements to the validity coefficients for this dimension using the Empirical and Merged keys, in particular. However, some serious questions remain as to the cross-sample generalizability of the positive findings reported in Study 2.

Specifically, to what degree are the reported validities spuriously inflated because of capitalization on chance, resulting in statistically significant correlations simply on the basis of chance, or Type I error. Whenever items from a larger set are chosen on the basis of their relation to the criterion, it is inevitable that some of the newly-selected items, when examined in a new sample, will not be related to the criterion. We would expect that the observed validity coefficients (especially those resulting from the Empirical key) are not the true validities, and so we would expect to see shrinkage in the cross-validated coefficients. This expectation comes from two primary sources. First, it will be recalled that reduced-item subsets were selected on the basis of maximal item validities for the original sample. Second, in the derivation of the Empirical and Merged keys, a total of six hundred correlations between the action elements and the Overall Performance Criterion were considered in order to determine the final Quality of Judgement value assigned each action element. Clearly, the large number of correlations required for Study 2 increased the likelihood that chance factors could operate with so many correlations used in this process of alternate
scoring key development. A final consideration is that, although action elements endorsed by less than three participants across the sample were not used in the correlational analyses, this cut-off was somewhat arbitrary, and low endorsement rates for other action elements may not yield stable correlations.

These considerations pointed, then, to the likelihood that shrinkage would occur in a cross-sample validation. It should be noted, however, that using a fairly large initial sample (321 subjects) will tend to reduce shrinkage because of smaller sampling errors. At this point, it is difficult to determine the extent of the phenomenon of capitalization on chance in order to more accurately assess the true validity of the Quality of Judgement dimension as measured by the Empirical and Merged keys. It is possible, nonetheless, to estimate (albeit roughly) the degree of possible capitalization on chance by considering the number of statistically significant items and the number that would be expected from chance or Type I error. The most accurate means, however, to directly quantify the effect of this artifact is to apply the same method (multiple key derivations and reduced-item subsets) to a new sample.

It should be noted that one problem with this approach is the introduction of possible bias and variation in the results due to differences in the sample (geographic, cultural etc.). A double cross-validation, in which the original sample is split and evaluated against itself, would be an effective way to estimate the true validity. However, a substantially larger sample would be needed than that available here. Given this limitation (of a changing sample), the most telling method to determine the degree to which the promising results observed in Study 2 are true, as opposed to inflated, predictive validities of the in-basket Quality of Judgement scale would be to re-administer the instrument in another setting, applying the same three scoring keys to the same three combinations of item sets.
RATIONALE AND HYPOTHESES FOR THE PRESENT STUDY

General Summary

The Literature Review has provided a considerable number of summaries of the empirical findings from forty years of research on the in-basket exercise. Consequently, specific summary observations, such as the marginal evidence of reliability and mixed evidence of criterion-related validity, will not be restated here. Instead, the broader implications from these observations will be incorporated with salient findings from Preliminary Studies 1 and 2 in order to identify the objectives and to develop the hypotheses for the present research study. (It should be noted that, although there is agreement among researchers for the need for further construct validity examinations of the in-basket, this area will not be addressed in the present study.)

Summaries from the literature and past research findings, especially those research results reported in the two preliminary studies, point to particular areas needing further investigation. For example, it was observed in the Literature Review that, given its widespread use, there was a relative lack of sound empirical validation for the in-basket exercise. A need for an approach to in-basket research and development which considered both psychometric and practical implications was identified. The past dilemma of developing an objectively-scored in-basket with maximal reliability and criterion validity which could also be readily scored by industry personnel appears to remain after nearly forty years of research on this popular instrument. It was recognized that research efforts which focused on methods of scoring the instrument seemed a viable way to simultaneously address the somewhat contrary needs of empirical soundness and feasibility or practicality in application.
These considerations and concerns guided the focus and goals of the two preliminary studies reported earlier.

The results of the first preliminary study examining a newly-revised in-basket, the TSIB, suggested that, although the traditional, panel-based scoring system yielded reliable dimensions, the criterion-related validity results were very modest. Moreover, an in-basket exercise which required nearly as long to score as it did to administer (1 1/2 hours) would simply not be practical in industry. The economic pressures resulting from the current recession and the high costs associated with lengthy scoring- and training-time would likely pose significant impediments to organizations considering the large-scale application of the in-basket in its standard form (i.e., written, free-format responses).

The second preliminary study was designed to investigate the relative psychometric and practical benefits of less conventional, more empirical, in-basket scoring strategies. The call from Brannick et. al (1989) for further research into in-basket scoring systems has therefore been answered, in part, by Study 2. This study also examined the effects of expanding the Scoring Manual to include more scorable units of action by the re-analysis of the Unusual actions, the effects of selecting several subsets of items for scoring, and the effects of developing and applying several Quality of Judgement scoring keys to these reduced-item subsets. As noted, scoring key modifications and a reduced-item scoring approach did not result in a reduction in reliability and validity. In fact, significantly higher criterion-related validity coefficients across a greater number of dimensions were observed. Furthermore, substantial reductions in scoring time (and training time required for scoring) were realized. Although promising, the results from this most recent investigation must be researched further before the conclusion can be made with confidence that the observed
validity coefficients are, in fact, more accurate estimates of the true validity for this critical dimension of administrative performance.

Proposal and Hypotheses

Accordingly, in the present study, a cross-validation of the new Quality of Judgement scoring approaches in another, related setting will be conducted. The same set of three scoring keys, combined with the reduced-item scoring approach (using the same item-subsets) will be applied to assess the extent of shrinkage in the cross-validated validity coefficients. It is hypothesized that detectable, but not substantial, reductions in the Quality of Judgement validity coefficients will be observed. It is further hypothesized that a new application of the reduced-item scoring approach will reaffirm its ability to provide optimal psychometric properties while also addressing more practical, application-oriented concerns. It should be made clear that, because scoring the full 21-item set no longer appears necessary or feasible, only those subsets of items used in Study 2 will be re-analyzed (8-, 10-, and 12-items). The full set of 21 items will not be scored for the new sample.

In addition to a cross-validation of the Quality of Judgement scoring strategies introduced in Study 2, further research and continued development of the TSIB in several additional areas appears warranted. In particular, ways to further refine and understand the remaining dimensions are needed.

The seven in-basket dimensions that comprise the TSIB can be grouped into the two broad categories of: performance and stylistic dimensions. The two performance dimensions, Productivity and Quality of Judgement, generally reflect more evaluative, maximum-performance aspects of in-basket responses, whereas the remaining stylistic dimensions are frequency-based descriptive measures of the manner in which a
participant tends to respond. To summarize, the dimensions measured in the TSIB were as follows:

**Stylistic:**
1. Planning and Organizing
2. Interpersonal Relations
3. Leadership in a Supervisory Role
4. Managing Personnel
5. Analysis and Synthesis in Decision-Making

**Performance:**
6. Productivity
7. Quality of Judgement

**Criterion measurement**

A company performance appraisal instrument, administered annually, will be used as the criterion measure for the present study. Data from 15 dimensions of work-related performance, measured by a 7-category ordinal scale involving descriptors (rather than numbers) reflecting levels of performance, will be collected from the Human Resources Department records for the previous two years of participants' performance (1989 and 1990). Numerical values will be assigned to each category (e.g., "excelling" = 1, "unsatisfactory" = 7). Averages will then be computed for each dimension across the two sets of ratings.

**Quality of Judgement Scores**

Predictive validities of scores based on the Panel, Empirical and Merged scoring keys. The critical Quality of Judgement dimension will form the basis for the cross-validation of the present study with a comparative analysis of the predictive accuracy of
the Quality of Judgement scores as measured by the three scoring keys (Panel 2--Conservative, Empirical, and Merged keys), and as applied to each of the three TSIB reduced-item sets in the two studies. The focus, then, will be to examine the effects of capitalization on chance by re-applying the same scoring keys used in Study 2 to measure the Quality of Judgement dimension in a new, second sample. The main research question to be addressed is to what degree the validity coefficients reported in Study 2 for the 8-, 10-, and 12-item subsets will hold up when applied in a new setting. It is hypothesized that the Merged and, to a lesser degree, the Empirical key used to measure the Quality of Judgement dimension will result in the largest validities for the three scoring keys.

Additional analyses based on the Panel scoring key. Whereas Preliminary Studies 1 and 2 both used a sample from Company 1, the present study will use a new sample from Company 2 which will allow, through sample comparisons, a more finely-tuned analysis of the Panel scoring keys applied to the Quality of Judgement dimension. For example, the question of whether the weak link in Study 1 arose from the specific judgements (i.e., calibre of decisions) from Panel 1, or from the actual method of combining the judgements of the panel can now be examined by separating the two critical variables: origin of the panel (Panel 1 versus Panel 2) and the method of combining panel judgements (Liberal or majority method with a three point rating scale used in Study 1 versus Conservative method using a 5 point rating scale used in Study 2) and examining several combinations of these variables in relation to the two datasets (Company 1 versus Company 2). A framework for the planned analyses is illustrated in Figure 3. The shaded areas represent analyses already conducted in Studies 1 and 2, namely Study 1 involved the Liberal Method using Panel 1 applied to the Company 1 dataset and Study 2 involved the Conservative Method using Panel 2 applied to the Company 1 dataset.
Figure 5. Design grid for planned panel analyses in terms of origin of the Panel key, of method of combining panel judgements, and of the dataset to which the keys will be applied.
In addition to the two analyses already conducted, then, six additional analyses will also be carried out:

1. Conservative combination method using Panel 1 applied to Company 1 dataset.

2. Conservative combination method using Panel 1 applied to Company 2 dataset.


4. Liberal combination method using Panel 1 applied to Company 2 dataset.

5. Liberal combination method using Panel 2 applied to Company 1 dataset.


It is hypothesized that, because of the greater consensus required among the panel members and the use of a 5-point, rather than 3-point, scale, the Conservative method of combining judgements will result in greater validity coefficients than the Liberal method. Also, it is expected that, in general, Panel 2 ratings will yield higher validities than Panel 1, because of the expanded list of action elements and the larger panel size used for the action element ratings. Lastly, "intra-company" analyses, comparing results wherein the source of the panel and the dataset match, (e.g., Panel 1 applied to Company 1 dataset; Panel 2 applied to Company 2 dataset) will produce greater validity coefficients than "inter-company" analyses (Panel 1 applied to Company 2 dataset). This pattern is expected because intra-company analyses should tend to eliminate possible confounding effects due to differences in corporate culture or geographical location.
Additional Performance Dimensions in the TSIB

Understanding of Situation Questionnaire. In the present study, a new eighth dimension (performance rather than stylistic) will be added in order to measure the participant's understanding of the issues, problems and broader implications of the scenario. Although it is often clear what actions the participant took on the exercise, it is often not known why a particular action was chosen. It is hypothesized that more information about a participant's rationale or thinking will constitute useful predictive information about that person's administrative potential. As suggested by the fairly low internal consistency results in Study 2, it may be that Quality of Judgement is made up of multidimensional, independent aspects of appropriateness of actions. Thus, by asking participants additional questions about their attitudes, strategies, and rationales behind the actions they chose, it is believed more valuable, predictive information will be obtained from participants' particular approaches to problems in the exercise. Although such data is perhaps more directly cognitive than is generally true of in-basket data, it nonetheless seems capable of improving in-basket assessment. To this end, a 33-item multiple choice Understanding of Situation Questionnaire will be designed and administered upon completion of the in-basket exercise. A sample item from the TSIB Understanding of Situation Questionnaire is provided in Figure 6.

The development of the scoring key for the Understanding of Situation Questionnaire will involve the same 21-member panel that provided the New Panel Quality of Judgement ratings. They also completed this questionnaire and the pattern of their selections will be analyzed and a preliminary key using a 0, 1, 2 scoring system assigned each of the five options per question will be applied. The particular value
At the moment, probably the most important issue to be dealt with is:

a. the budget overrun on overtime
b. conflicts among employees
c. public image/customer satisfaction
d. low productivity
e. low charitable donations

Figure 6. Sample Item from the Understanding of Situation Questionnaire.
assigned to each option will depend on the level of endorsement from the expert panel. Across the five response options provided for an item, the panel members' pattern of option selection will be examined and an empirical, followed by a logical approach, will be used to assign scoring weights. The empirical approach is based on the frequencies of option selections within an item. If, for example, option "b" were selected most frequently by panel members, it would initially be keyed "2" whereas if option "d" were chosen by very few members (or none), it would be keyed "0". Options receiving moderate endorsement by the panel (i.e., options neither maximally nor minimally selected) would initially be keyed "1". Following this empirically-based assignment of scoring weights, final determination of the scoring key will be made by a logical, or rational consideration of the relative differences between option selection rates. If, for example, the two most frequently selected options within an item differed by one or perhaps two panel members' selection, both options may be keyed "2". Similarly, if two moderately-endorsed options showed a minimal difference in their endorsement rates, they may both be keyed "1".

The results from this instrument will be analyzed by correlating both the individual item scores and overall questionnaire scores with the company performance appraisal ratings. Item-analyses using item-reliability and item-validity indices will be conducted in order to select a subset of items from the questionnaire with maximal reliability and validity. An important advantage of the questionnaire format compared to the standard ETS Reasons for Action form or a post-exercise interview is that the multiple-choice scoring format requires less administration and scoring-time and it provides improved reliability by using a quantitative, objectively-applied scoring key.

**Productivity.** Several different ways of operationalizing Productivity to increase the validity of this dimension will be examined. As seen in Study 2, the
Productivity dimension did not contribute significantly to in-basket performance for men or women. It will be recalled that, in Study 1, Productivity was operationalized as the total number of actions taken across the exercise (including Unusual actions), as well as the number of items attempted.

In the present study, several specific methods of defining this dimension will be used. In essence, the approach will be based on that described in Hakstian et al. (1986), involving the aggregation of smaller, more concrete units of output. In all, seven units of productivity will be assessed: (a) the number of items attempted across 21 items, (b) the number of letters or memos written across 21 items, (c) a score reflecting the number of words written per memo or letter across the 21 items will be determined, (d) number of actions scheduled for a definite time, (e) number of entries made on the calendar, (f) whether a "things to do" list was completed, and (g) whether a "summary" list of the items was completed. Productivity linear composites will then be derived to determine the optimal Productivity operationalization by correlating the composite scores with company performance appraisal ratings.

**In-Basket Stylistic Dimensions**

The efficacy of logically assigning the Stylistic dimensions will be re-examined using the 8-, 10-, and 12-item subsets. Using the expanded list of action elements described in Study 2, the total scores for each stylistic dimension (derived by summing the number of action elements logically coded for each dimension) will be correlated with company performance appraisal ratings. The purpose for this analysis is examine the validity coefficients for these reduced-item dimensions when applied in the Company 2 dataset used in the present study. The intercorrelations of the stylistic dimensions will also be calculated, separately be gender, with the expectation that, as in Study 1, a fairly high level of intercorrelation among the stylistic dimensions will be
seen. Results can then be compared with those obtained in Study 2 using the Company 1 dataset (although broadly because of the use of the expanded list of action elements in the present study).

A factor analysis of the action element endorsements will also be conducted because of the high intercorrelations among the dimension scores seen in Study 2 (Tables 6 and 7). This analysis will produce an empirical, as opposed to rational, set of administrative dimensions of performance from the second dataset. Using factor loadings, each action element will be coded for the most relevant factor. As observed in the Literature Review, there is some disagreement regarding the complexity of in-basket performance dimensions (Frederiksen, 1966; Lopez, 1966). It is hypothesized that this analysis will help resolve the conflicting results and conclusions seen to date. Criterion-related validities of each factorially-derived dimension will be also determined by correlating dimension scores with the company performance appraisal ratings.

Additional New Measures

In addition to the introduction of the new Understanding of Situation dimension and a re-examination of the existing dimensions as outlined above, two new measures will also be investigated.

**Number of High Priority Items attempted.** The first new measure involves an evaluation, by the 21-member panel of Study 2, of the priority or urgency which should be afforded each item. The Priority rating was based on a consideration of the immediacy of action required by the item. Each panel member independently assigned a priority rating to each item using a 3-point scale in which "1" indicated a low priority item where action could be deferred one calendar week or more. A weight of "2" stood for a medium priority item where action, although not necessarily required
immediately, is required within the week, and "3" was assigned to high priority items which required immediate action. Final identification of high priority items was made by considering the majority judgements of the panel.

In the present research, participants will receive one point for each high priority item attempted in the exercise, based on an evaluation of the entire set of 21 items, rather than a reduced-item subset. The total number of points, or high priority items attempted, will then be correlated with the company performance appraisal ratings. It is hypothesized that the number of high priority items attempted across the TSIB will show positive, significant validity when correlated with the criterion rating.

Scorer's Impression of Involvement. The second new measure follows from several promising results reported in the literature with the impressionistic Scorer's Rating of overall performance (e.g., Meyer, 1970). The predictive accuracy of a similar subjective measure will be explored in the present study. As a way to operationalize Lopez's (1966) crucial concept of "ego-involvement," a 3-point rating scale of the degree of involvement will be used by scorers to measure their impression of "no involvement" (assigned a 1) to "extreme involvement" (assigned a 3). The assigned Scorer's Impression of Involvement values will be correlated with the company performance appraisal ratings to determine whether it will contribute useful predictive information about participants' administrative performance.
METHOD

Participants and Setting

The participants were 321 employees of a large western Canadian utility company (176 males and 145 females). Sixty one percent of the sample were first-level managers (195 employees), while the remaining 126 participants were second-level managers. In the summer of 1991, participants were given the TSIB and a low-fidelity simulation in a concurrent validation study of these instruments.

To recruit participants, the Human Resources Department first generated a random sample of 1,000 first and second-level managers from a pool of approximately 2,000 employees across all major departments. Next, two factors guided the quasi-random selection of employees for possible participation: first, an equal gender balance of participants was sought, and secondly, equal employee involvement from across the seven main divisions of the company was desired. Potential participants were contacted by company mail to solicit their voluntary involvement in the study. Participants were provided with extensive, confidential feedback on their exercise results.

The two instruments required three hours to complete. As described in Preliminary Study 1, participants were first given 1 1/2 hours to complete the TSIB, preceded by 15 minutes of scripted instructions. Upon completion of the exercise, participants were then given the Understanding of Situation questionnaire, for which no time limit was imposed. (The average time required to complete the post-exercise questionnaire was 30 minutes.) Participants were then administered the second, untimed, low-fidelity instrument, which usually required 1 hour to complete. Typically, two administration sessions were held each testing day, and participants
were given the option of attending a morning or afternoon session. In total, 30 testing sessions were conducted over a testing period of approximately six weeks, in groups ranging from 5-6 employees to 20-25 employees at a time.

Assessment Measures

Predictor measures.

1. The Telephone Supervisor In-Basket Exercise (TSIB).

21 items. The TSIB used was identical to that employed in Studies 1 and 2, except for the inclusion of the Understanding of Situation Questionnaire (33 items).

2. The Supervisory Profile Inventory (SPI).

The SPI is a low-fidelity simulation consisting of two parts: a) Part A, a biodata questionnaire designed to measure employees' personal interests, background experiences, and opinions (50 items) and, b) Part B, also a questionnaire format, which assessed individual differences in managerial style (22 items). Because the SPI is not a focus of the present work, it shall not be discussed further.

Criterion measures.


Initially, a company performance appraisal instrument, administered annually, was selected as the criterion measure for the present study. Fifteen dimensions of work-related performance along a 7-point scale were measured. Criterion data were collected for each participant along the 15 performance
dimensions measured for both 1989 and 1990. However, the performance ratings were severely positively skewed (the low end of the scale measured better performance whereas the high end of the scale measured poorer performance) with little variability. Consequently, the company appraisal instrument appeared to be demonstrating little discrimination in managerial performance between participants. The likely causes of the limited criterion variability were the cumulative effects from several judgemental rating biases such as leniency, central tendency, and halo bias, as well as missing data from incomplete files.

2. The Employee Appraisal Inventory (EAI).

Because of these concerns of low variability in the company performance appraisal measure, an additional questionnaire, the Employee Appraisal Inventory (EAI), was then administered by company mail to the participants' immediate supervisors in order to collect further criterion data on work-related performance. All participants were informed by company mail of the need for this additional criterion measure and were assured that EAI results were confidential and for research (validation) purposes only. Participants could refuse to allow the collection of this additional criterion data with impunity. The EAI, which required approximately 10 minutes to complete, was returned for 296 of the 321 participants (159 males; 137 females). Of the 25 non-completions, five participants refused the EAI while the remaining 20 inventories were not returned by the immediate supervisor.

The EAI is made up of the same set of three behavioural observation scales (BOS) used to measure each of 12 performance dimensions previously described in Studies 1 and 2 in the Preliminary Studies section. As noted
earlier, the development of the BOS scales is described more fully in an article by Hakstian et al. (1991). The instrument requires the supervisor to rate the frequency of certain job behaviours (corresponding to the performance dimensions) observed in the employee. The order of the 36 BOS statements, each measured by a scale ranging from 1 (low) to 5 (high), was rotated both by dimension and by tone (i.e., a positively-worded statement was followed by a negatively-worded one). An example of one of the three BOS scale statements used to measure the management performance dimension Planning/Organizing/Control is the following:

Establishes a plan for the fiscal year; maps out when each event must take place in order to meet the stated goals; allows for unexpected circumstances.

Almost Never 1 2 3 4 5 Almost Always

A factor analysis of the 12 performance dimensions conducted in Study 2 resulted in four global management performance dimensions: (a) Interpersonal Effectiveness, made up of the dimensions Leadership, Behaviour Flexibility, and Sensitivity; (b) Methodical/Stable Performance, made up of the dimensions Planning/Organizing, Work Ethic, Initiative, Performance Stability; (c) Insightful/Decisive Performance, made up of dimensions Analysis, Judgement and Decisiveness; and (d) Communication Effectiveness, made up of Oral Communication and Written Communication. It will be recalled that, in the preliminary studies, three BARS scales and three BOS scales were constructed for each performance dimension. Whereas the preliminary studies based each of the four global management dimensions on a sum of the BARS and BOS
scales, the present study, in order to facilitate the acquisition of criterion data, used only the BOS scales in the measurement of the performance dimensions.

These four dimensions were then added to yield a fifth, more molar criterion called Overall Management Performance (OMP). It is this Overall outcome measure that figures most prominently in the criterion-related analyses of the present study. It should be made clear that the four global dimensions and the OMP to be used here were based on half the number of scales used in earlier studies because of the exclusion of the BARS scales. (Where possible and appropriate, comparable criterion correlations between the OMP and relevant TSIB dimensions from Study 2 will be reported, in order to facilitate the evaluation of the scoring keys used in the Quality of Judgement dimension cross-validation.)
RESULTS

Overview of Study Design and Data Analysis

Design

The primary focus in the present study was to assess the degree of shrinkage in the validity coefficients of a cross-sample validation of Quality of Judgement scores derived from the three scoring keys (Panel 2--Conservative, Empirical, and Merged keys) previously applied in Study 2 to the three reduced-item subsets (8-, 10-, and 12-items). Accordingly, a concurrent validation of the TSIB was undertaken by administering it to a second, related sample. The validities of the Quality of Judgement scores were examined and the validity coefficients from the two applications were compared. It should be clear that the main difference in the version of the TSIB used in Preliminary Study 2 compared to that used in the present study is the introduction of a multiple-choice questionnaire administered upon completion of the TSIB.

The concurrent validation allowed further specific analyses, involving new measures (Number of Priority Items Attempted and Scorer's Impression of Involvement), new operationalizations of existing dimensions (e.g., Productivity) and a factor-analytic derivation of TSIB stylistic dimensions to be carried out. These analyses are summarized in the following section.

Data Analysis

Reliability. In order to validate the TSIB, data from existing measures of job performance were collected from the company (against which TSIB scores were correlated). The reliability (alpha coefficient) of the criterion performance dimensions was calculated. Reliability estimates (internal consistency) of the TSIB Quality of
Judgement dimension as measured by the three scoring keys, applied to the three reduced-item subsets, were also determined. In addition, internal consistency estimates of the reliability of the TSIB stylistic and remaining performance dimensions (Understanding of Situation Questionnaire and Productivity) were carried out.

**Criterion-related validity.** The validity of the Quality of Judgement scores, derived from the three scoring keys and applied to the three reduced-item subsets, was examined by correlating dimension scores with on-the-job measures of performance. Comparisons with previously-obtained validity coefficients for the same reduced-item subsets, scored using the same scoring keys, were made.

In addition, to further assess the predictive accuracy of the Panel-based derivation of the Quality of Judgement scoring key, Quality of Judgement scores derived in the following six ways were correlated with job performance measures: (a) Conservative panel judgement combination method using Panel 1 applied to Company 1 dataset, (b) Conservative combination method using Panel 1 applied to Company 2 dataset, (c) Conservative combination method using Panel 2 applied to Company 2 dataset, (d) Liberal combination method using Panel 1 applied to Company 2 dataset, (e) Liberal combination method using Panel 2 applied to Company 1 dataset, and (f) Liberal combination method using Panel 2 applied to Company 2 dataset. Thus, three aspects of the Panel-based Quality of Judgement keys—the method of combining panel judgements, the origins of the Panel key, and the origins of the data to which the keys will be applied—were targeted in order to identify those factors related to the Panel key that are most useful in predicting administrative ability.

The TSIB stylistic dimensions (Planning and Organizing Work to Analysis and Synthesis in Decision-Making) were measured in the present study using the expanded list of action elements applied to the reduced-item subsets. Criterion-related validity
results and the intercorrelations of the stylistic dimensions were then compared with those obtained in Study 2 (which used the Company 1 dataset). A factor analysis of the action element endorsements from the Company 2 dataset was conducted in order to produce an empirical, as opposed to rational, set of descriptive dimensions of administrative performance. Reliability estimates (internal consistency) and criterion-related validities of each factorially-derived dimension were determined.

Validity coefficients were corrected for both criterion unreliability and range restriction (Schmidt & Hunter, 1981). The correction for range restriction was based on logic more fully described in Hakstian et al (1991). Briefly, as a maximum performance measure, administrative skill assessment relied on an unrestricted variance estimate that was midway between the general population variances published in standardized cognitive measure test manuals and the restricted variance estimates supplied by the data at hand.

A brief comment on the Results. It should be noted that, in the ensuing presentation of results, the data are merely described and briefly summarized. More detailed analyses of the hypotheses and consideration of the implications of the results are found in the following Discussion section.

Major Reliability and Criterion-related Validity Findings

Reliability of the Performance Appraisal Dimension Scores

A reliability analysis (Cronbach’s alpha) was conducted on the four global and the fifth more molar management criterion (previously described in the Methods section). These reliability estimates are reported in Table 8. The alpha coefficients of the four global management performance dimensions range from a low of .48 to a high of .92, with a mean of .80. The mean alpha coefficient of the four global management
dimensions increases to .84 when the lowest, anomalous result (.48 for female Communication Effectiveness) is omitted. For comparison, a reliability estimate based on a generalizability analysis of the four global performance dimensions measured in Preliminary Study 1 (involving Company 1) and then corrected for length for application in the present study (the number of scales used here were half those of Study 1) yielded single-facet generalizability (alpha) coefficients ranging from .83 to .90, with a mean of .87. Thus, the four global management performance dimensions show consistent evidence of strong reliability. In addition, as shown in Table 8, the fifth most molar performance criterion, Overall Management Performance or OMP, is highly reliable.

Cross-Validation of the Quality of Judgement Scores

Of central interest in the present study is the degree to which the three TSIB Quality of Judgement keys (Panel 2--Conservative, Empirical, and Merged), as applied to the three reduced-item subsets (8-, 10-, and 12-item) yield similar criterion-related validity results when applied to the Company 2, rather than Company 1, dataset. Table 9 presents the resulting validity coefficients, with relevant coefficients from the previous application (correlated against the same OMP criterion) provided in parentheses beside each result. As expected, the Merged key yields higher validity coefficients than the Empirical key, from which only two of six correlations (the 8- and 10-item subsets for women) are significant. As noted earlier, these results for the Quality of Judgement scores will be reviewed in greater depth in the following Discussion section. (The reliability of the Panel 2--Conservative key, selected because it yielded the highest overall validity coefficients, was determined and the findings are shown later, in Table 18, along with the reliability of the remaining TSIB dimensions.)
Table 8

Reliability Analysis (Alpha Coefficient) of the Five Performance Appraisal Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpersonal Effectiveness</td>
<td>.81</td>
<td>.76</td>
</tr>
<tr>
<td>Methodical/Stable Performance</td>
<td>.92</td>
<td>.86</td>
</tr>
<tr>
<td>Insightful/Decisive Performance</td>
<td>.91</td>
<td>.85</td>
</tr>
<tr>
<td>Communication Effectiveness</td>
<td>.79</td>
<td>.48</td>
</tr>
<tr>
<td>Overall Management Performance</td>
<td>.96</td>
<td>.92</td>
</tr>
</tbody>
</table>
Table 9

Bivariate Correlations between Overall Management Performance and Quality of Judgement Scores Obtained from Three Scoring Keys and from Three Item Sets

<table>
<thead>
<tr>
<th>Item Sets</th>
<th>Scoring Key</th>
<th>Panel 2</th>
<th>Empirical</th>
<th>Merged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(Conservative)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td></td>
<td>.30** (.29)</td>
<td>(.61)</td>
<td>.19* (.60)</td>
</tr>
<tr>
<td>10 item set</td>
<td></td>
<td>.26** (.23)</td>
<td>(.64)</td>
<td>.18* (.62)</td>
</tr>
<tr>
<td>12 item set</td>
<td></td>
<td>.26** (.24)</td>
<td>(.65)</td>
<td>.18* (.61)</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 item set</td>
<td></td>
<td>.25** (.28)</td>
<td>.19* (.63)</td>
<td>.23* (.59)</td>
</tr>
<tr>
<td>10 item set</td>
<td></td>
<td>.32** (.26)</td>
<td>.19* (.62)</td>
<td>.23* (.62)</td>
</tr>
<tr>
<td>12 item set</td>
<td></td>
<td>.25** (.23)</td>
<td>(.64)</td>
<td>.23* (.63)</td>
</tr>
</tbody>
</table>

Note. Reported $r$'s have been corrected for effects of range restriction and criterion unreliability. Values in parentheses are bivariate correlations between Overall Management Performance and Quality of Judgement scores obtained in Study 2, provided to facilitate comparison between the previous and present study.

* $p < .05$ (one-tailed). ** $p < .01$ (one-tailed).
As noted in the Rationale and Hypotheses for the Present Study section, the application of Quality of Judgement scores to a new dataset allowed more finely-grained analyses of the Panel-based scoring key. Table 10 shows the results of six separate, but related, analyses based on varying three factors: the origin of the Panel key, the method of combining panel judgements, and the dataset to which the keys were applied. Once more, these results and their relation to the hypotheses outlined earlier will be reviewed in the following Discussion section.

Other Performance Dimensions in the TSIB

Understanding of Situation Questionnaire. As described in the Rationale and Hypotheses for the Present Study section, the Understanding of Situation Questionnaire consisted of 33 items, with five response options per item, designed to measure the participant’s understanding of the issues, problems and broader implications of the scenario presented in the in-basket exercise. The panel whose ratings provided the scoring key (previously described) strongly recommended the deletion of one particular item, which was subsequently eliminated to yield a final set of 32 items comprising the questionnaire.

Table 11 presents the reliability (internal consistency) and criterion-related validity results of the 32-item questionnaire. The resulting validity coefficient is modest for females but marginal for males. Therefore, item analyses were conducted in order to select a subset of items from the original 32 which, across both genders, would yield stronger validity coefficients for both groups. Item-reliability and item-validity indices were first calculated and then represented on a graph to better select those items with maximal reliability and validity for both genders. Several sets of items were examined, with the inclusion of individual items based more on their contribution to validity, than to reliability. Table 12 presents the reliability and
Table 10

**Bivariate Correlations between Overall Management Performance and Quality of Judgement Scores Obtained Using Scoring Keys based on Different Methods of Combining Panel Judgements (Liberal or Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), and Applied to Different Datasets (Company 1 or Company 2).**

<table>
<thead>
<tr>
<th>Item Sets</th>
<th>Quality of Judgement Scoring Approach</th>
<th>8-item</th>
<th>10-item</th>
<th>12-item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Conservative Method, Panel 1, Company 1 dataset</td>
<td>.30**</td>
<td>.24**</td>
<td>.25**</td>
</tr>
<tr>
<td>2.</td>
<td>Conservative Method, Panel 1, Company 2 dataset</td>
<td>.19*</td>
<td>.18*</td>
<td>.21*</td>
</tr>
<tr>
<td>3.</td>
<td>Conservative Method, Panel 2, Company 1 dataset</td>
<td>.29**</td>
<td>.23**</td>
<td>.24**</td>
</tr>
<tr>
<td>4.</td>
<td>Conservative Method, Panel 2, Company 2 dataset</td>
<td>.30**</td>
<td>.26**</td>
<td>.26**</td>
</tr>
<tr>
<td>5.</td>
<td>Liberal Method, Panel 1, Company 1 dataset</td>
<td>.37**</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6.</td>
<td>Liberal Method, Panel 1, Company 2 dataset</td>
<td>.26**</td>
<td>.23**</td>
<td>.18**</td>
</tr>
<tr>
<td>7.</td>
<td>Liberal Method, Panel 2, Company 1 dataset</td>
<td>.32**</td>
<td>.25**</td>
<td>.25**</td>
</tr>
<tr>
<td>8.</td>
<td>Liberal Method, Panel 2, Company 2 dataset</td>
<td>.29**</td>
<td>.23**</td>
<td>.24**</td>
</tr>
</tbody>
</table>
Table 10 (cont.)

Bivariate Correlations between Overall Management Performance and Quality of Judgement Scores Obtained Using Scoring Keys based on Different Methods of Combining Panel Judgements (Liberal or Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), and Applied to Different Datasets (Company 1 or Company 2).

<table>
<thead>
<tr>
<th>Item Sets</th>
<th>Quality of Judgement Scoring Approach</th>
<th>8-item</th>
<th>10-item</th>
<th>12-item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Conservative Method, Panel 1, Company 1 dataset</td>
<td>.19*</td>
<td>.20*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Conservative Method, Panel 1, Company 2 dataset</td>
<td>.37**</td>
<td>.38**</td>
<td>.36**</td>
<td></td>
</tr>
<tr>
<td>3. Conservative Method, Panel 2, Company 1 dataset</td>
<td>.28**</td>
<td>.26**</td>
<td>.23**</td>
<td></td>
</tr>
<tr>
<td>4. Conservative Method, Panel 2, Company 2 dataset</td>
<td>.25**</td>
<td>.32**</td>
<td>.25**</td>
<td></td>
</tr>
<tr>
<td>5. Liberal Method, Panel 1, Company 1 dataset</td>
<td>.29**</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6. Liberal Method, Panel 1, Company 2 dataset</td>
<td>.35**</td>
<td>.35**</td>
<td>.36**</td>
<td></td>
</tr>
<tr>
<td>7. Liberal Method, Panel 2, Company 1 dataset</td>
<td>.26**</td>
<td>.25**</td>
<td>.23**</td>
<td></td>
</tr>
<tr>
<td>8. Liberal Method, Panel 2, Company 2 dataset</td>
<td>.30**</td>
<td>.30**</td>
<td>.32**</td>
<td></td>
</tr>
</tbody>
</table>

Note. Reported $r$'s have been corrected for the effects of range restriction and criterion unreliability. Correlations for the 10- and 12-item sets, scored by the Liberal Method, Panel 1 key, applied to the Company 1 dataset are not available. All non-significant correlations were omitted.

*p < .05 (one-tailed). **p < .01 (one-tailed).
Table 11

Validity Coefficients between Overall Management Performance (OMP) and Understanding of Situation Questionnaire Total Score (32-item): Alpha Reliability Estimates for Total Score

<table>
<thead>
<tr>
<th></th>
<th>Bivariate Correlation with OMP&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Understanding of Situation</td>
<td>.19&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.29&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Questionnaire Total Score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Alpha Coefficient</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Understanding of Situation</td>
<td>.24</td>
<td>.38</td>
</tr>
<tr>
<td>Questionnaire Total Score</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.  
<sup>a</sup> The reported r's have been corrected for the effects of range restriction and criterion unreliability.  
<sup>b</sup> Significant at the .05 level (one-tailed test).  
<sup>c</sup> Significant at the .01 level (one-tailed test).
Table 12

Validity Coefficients between Overall Management Performance and Understanding of Situation Questionnaire Total Score (20-item) and Reliability Estimates for Total Score

<table>
<thead>
<tr>
<th></th>
<th>Bivariate Correlation with OMP&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Alpha Coefficient</th>
<th>Split Half</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
</tr>
<tr>
<td>Understanding of Situation</td>
<td>.35</td>
<td>.34</td>
<td>.23</td>
</tr>
<tr>
<td>Questionnaire Total Score</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. <sup>a</sup> The reported r's have been corrected for the effects of range restriction and criterion unreliability. Both are significant at the .001 level (one-tailed tests).
criterion-related validity results for the most promising subset of items, comprising 20 items selected for retention in the final version of the Understanding of Situation Questionnaire.

Productivity. In its original operationalization applied in Preliminary Studies 1 and 2, Productivity was measured by the sum of the total number of items attempted, the total number of action elements, and the total number of Unusual actions taken across the exercise. It will be recalled that, for both males and females, criterion-related validity preliminary study findings were non-significant.

In the present study, several new ways of operationalizing Productivity were derived, based on the aggregation of smaller, concrete units of output, in order to examine the effect on the criterion-related validity of this performance dimension. In total, seven units of productivity were assessed: (a) the number of items attempted across 21 items, (b) the number of letters or memos written across 21 items, (c) a score reflecting the number of words written per memo or letter across the 21 items, (d) number of actions scheduled for a definite time, (e) number of entries made on the calendar, (f) whether a "things to do" list was completed, and (g) whether a "summary" list of the items was completed. More standard indices of output were also calculated for each item-subset, comprised of the total number of action elements and total number of Unusual actions taken over the 8-, 10-, and 12-item subsets and are referred to as Total action (8), Total action (10), and Total action (12), respectively. Productivity linear composites were also derived in order to determine the optimal Productivity operationalization, based on the most promising of the individual units of output.
The upper portion of Table 13 presents the validity coefficients between the individual units of output and the OMP. The middle lower portion of the table displays the results of the correlations between the linear composites and the OMP, whereas the lower portion of Table 13 presents validity coefficients between the more standard operationalizations of Productivity (as an index of the overall quantity of participants' output) and the OMP.

**TSIB Stylistic Dimensions**

Using the expanded list of action elements prepared in Study 2, the total scores for each stylistic dimension (calculated by summing the number of actions elements logically coded for each dimension) were correlated with the OMP for each of the three item-subsets and are displayed in Table 14. As Table 14 shows, a distinct pattern across genders, but consistent within each gender is evident. Specifically, across the item-subsets, Interpersonal Relations and Leadership in a Supervisory Role are the only significant stylistic dimensions for men, whereas Planning and Organizing Work, Managing Personnel, and Analysis and Synthesis in Decision-Making yield useful predictive information for women.

Intercorrelations among the TSIB stylistic dimension scores were determined and are reported in Tables 15, 16, and 17 for the 8-, 10-, and 12-item subsets, respectively. Although more accurately considered a performance dimension, the Productivity dimension, measured in the same way as in Study 1, was included in these analyses for completeness. (The Quality of Judgement dimension was not included because of the variety of methods used to calculate scores.)
Table 13

Validity Coefficients between Overall Management Performance (OMP) and Productivity as Measured by Several Individual Units of Output, Linear Composites, and Standard Total Action Indices

<table>
<thead>
<tr>
<th>Productivity--Units of Output</th>
<th>Bivariate Correlation with OMP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
</tr>
<tr>
<td>Individual Units of Output</td>
<td></td>
</tr>
<tr>
<td>1. Item Total (across 21 items)</td>
<td></td>
</tr>
<tr>
<td>2. # Letters/memos (across 21 items)</td>
<td>.24**</td>
</tr>
<tr>
<td>3. Total # words written (across 21 items)</td>
<td>.24**</td>
</tr>
<tr>
<td>4. Definite actions</td>
<td></td>
</tr>
<tr>
<td>5. Calendar entries</td>
<td></td>
</tr>
<tr>
<td>6. &quot;Things to do&quot; list</td>
<td></td>
</tr>
<tr>
<td>7. &quot;Summary&quot; list</td>
<td></td>
</tr>
<tr>
<td>Linear Composites of Units of Output</td>
<td></td>
</tr>
<tr>
<td>1. Item Total, # Letters/memos, Total action (10)</td>
<td>.24**</td>
</tr>
<tr>
<td>2. Item Total, Total # words written, Total action (10)</td>
<td>.25**</td>
</tr>
<tr>
<td>3. Item Total, Definite actions, Total action (10)</td>
<td>.19*</td>
</tr>
<tr>
<td>4. Item Total (across 10 items), #Letters/memos, Total action (10)</td>
<td>.25**</td>
</tr>
<tr>
<td>Standard Approach to Scoring of Productivity</td>
<td></td>
</tr>
<tr>
<td>1. Total action (8-item)</td>
<td>.18*</td>
</tr>
<tr>
<td>2. Total action (10-item)</td>
<td>.18*</td>
</tr>
<tr>
<td>3. Total action (12-item)</td>
<td>.18*</td>
</tr>
</tbody>
</table>

Note. Reported r's have been corrected for the effects of range restriction and criterion unreliability. All non-significant correlations were omitted.

*p < .05 (one-tailed). **p < .01 (one-tailed).
Table 14

Validity Coefficients between Overall Management Performance (OMP) and TSIB Stylistic Dimension Scores

<table>
<thead>
<tr>
<th>Stylistic Dimension</th>
<th>Bivariate Correlation with OMP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
</tr>
<tr>
<td><strong>8-item set</strong></td>
<td></td>
</tr>
<tr>
<td>1. Planning and Organizing Work</td>
<td></td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td></td>
</tr>
<tr>
<td>3. Leadership in Supervisory Role</td>
<td></td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td></td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td></td>
</tr>
<tr>
<td><strong>10-item set</strong></td>
<td></td>
</tr>
<tr>
<td>1. Planning and Organizing Work</td>
<td></td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td></td>
</tr>
<tr>
<td>3. Leadership in Supervisory Role</td>
<td></td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td></td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td></td>
</tr>
</tbody>
</table>
Table 14 (cont.)

**Validity Coefficients between Overall Management Performance (OMP) and TSIB Stylistic Dimension Scores**

<table>
<thead>
<tr>
<th>Stylistic Dimension</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Planning and Organizing Work</td>
<td>.23*</td>
<td></td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td>.25**</td>
<td></td>
</tr>
<tr>
<td>3. Leadership in Supervisory Role</td>
<td>.26**</td>
<td>.18*</td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td></td>
<td>.21*</td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td></td>
<td>.25**</td>
</tr>
</tbody>
</table>

**Note.** All non-significant correlations were omitted.

*p < .05 (one-tailed). **p < .01 (one-tailed).
Table 15

Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores based on the 8-Item Set:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P &amp; O</td>
<td>IR</td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.69</td>
<td>.75</td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.47</td>
<td>.29</td>
</tr>
<tr>
<td>Managing Personnel</td>
<td></td>
<td>.28</td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 16

Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores based on the 10-Item Set:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>P &amp; O</th>
<th>IR</th>
<th>LSR</th>
<th>MP</th>
<th>A &amp; S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning and Organizing Work</td>
<td>.10</td>
<td>.11</td>
<td>.08</td>
<td>.64</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.76</td>
<td>.76</td>
<td>.13</td>
<td>.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.54</td>
<td>.18</td>
<td>.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>.17</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension</th>
<th>P &amp; O</th>
<th>IR</th>
<th>LSR</th>
<th>MP</th>
<th>A &amp; S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning and Organizing Work</td>
<td>.30</td>
<td>.28</td>
<td>.21</td>
<td>.69</td>
<td>.58</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.75</td>
<td>.76</td>
<td>.26</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.49</td>
<td>.28</td>
<td>.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>.27</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.72</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 17

Intercorrelations of TSIB Stylistic Dimensions and Productivity Scores based on the 12-Item Set:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Males</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P &amp; O</td>
<td>IR</td>
<td>LSR</td>
<td>MP</td>
<td>A &amp; S</td>
<td>P</td>
</tr>
<tr>
<td>Planning and Organizing Work</td>
<td>.13</td>
<td>.14</td>
<td>.12</td>
<td>.71</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.82</td>
<td>.79</td>
<td>.16</td>
<td>.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.61</td>
<td>.20</td>
<td>.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>.19</td>
<td>.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.77</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Females</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P &amp; O</td>
<td>IR</td>
<td>LSR</td>
<td>MP</td>
<td>A &amp; S</td>
<td>P</td>
</tr>
<tr>
<td>Planning and Organizing Work</td>
<td>.34</td>
<td>.34</td>
<td>.26</td>
<td>.77</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>.82</td>
<td>.81</td>
<td>.28</td>
<td>.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership in a Supervisory Role</td>
<td>.59</td>
<td>.31</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managing Personnel</td>
<td>.27</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis and Synthesis in Decision-Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Productivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reliability of the TSIB dimensions

Reliability estimates (split half and alpha coefficients) were calculated for the seven dimensions, with Productivity measured as the standard index of overall output (sum of total action elements and Unusuals) and Quality of Judgement measured by the Panel 2 (Conservative) key, for each of the three item-subsets. These results are provided in Table 18. Not surprisingly, a pattern of increasing reliability with larger item-subsets is seen.

Additional New Measures

Number of High Priority Items Attempted. Ratings of priority or urgency (based on a consideration of the immediacy of action required by the item) were provided by the 21-member panel of Study 2 using a 3-point scale where 1 indicated a low priority item and 3 indicated a high priority item. Final identification of High Priority items was made from the majority judgements of the panel. Consequently, six items in the TSIB were labelled as High Priority items. Participants received one point, along a 6-point scale, for each High Priority item attempted across the entire set of 21 items in the TSIB. This total "High Priority Score" was then correlated with the OMP. Table 19 presents the results from this analysis.

Scorer's Impression of Involvement. A 3-point rating scale of the degree of involvement was used by scorers to measure their subjective impression of participants' efforts on the TSIB. The level of participation and commitment in the actions taken were scored from a low of 1, indicating "no involvement", to 3, indicating "extreme involvement". No significant results, for either gender, emerge when the Scorer's Impression of Involvement scores is correlated with the OMP. It appears, then, that
## Table 18

**Reliability Estimates of TSIB Dimensions**

<table>
<thead>
<tr>
<th>TSIB Dimension</th>
<th>Alpha Coefficient</th>
<th>Split Half</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td><strong>8-item set</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Planning and Organizing Work</td>
<td>.16</td>
<td>.30</td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td>.33</td>
<td>.29</td>
</tr>
<tr>
<td>3. Leadership in Supervisory Role</td>
<td>.05</td>
<td>.35</td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td>.44</td>
<td>.42</td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td>.45</td>
<td>.42</td>
</tr>
<tr>
<td>6. Productivity</td>
<td>.45</td>
<td>.51</td>
</tr>
<tr>
<td>7. Quality of Judgement(^a)</td>
<td>.24</td>
<td>.37</td>
</tr>
</tbody>
</table>

| **10-item set**                                     |       |         |       |         |
| 1. Planning and Organizing Work                    | .38   | .48     | .51   | .48     |
| 2. Interpersonal Relations                          | .41   | .33     | .43   | .37     |
| 3. Leadership in Supervisory Role                   | .22   | .40     | .27   | .33     |
| 4. Managing Personnel                               | .43   | .42     | .48   | .45     |
| 5. Analysis and Synthesis in Decision-Making        | .59   | .58     | .66   | .68     |
| 6. Productivity                                     | .45   | .55     | .47   | .61     |
| 7. Quality of Judgement\(^a\)                        | .26   | .38     | .36   | .41     |
### Table 18 (cont.)

Reliability Estimates of TSIB Dimensions

<table>
<thead>
<tr>
<th>TSIB Dimension</th>
<th>Alpha Coefficient</th>
<th>Split Half</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>12-item set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Planning and Organizing Work</td>
<td>.45</td>
<td>.56</td>
</tr>
<tr>
<td>2. Interpersonal Relations</td>
<td>.47</td>
<td>.45</td>
</tr>
<tr>
<td>3. Leadership in Supervisory Role</td>
<td>.30</td>
<td>.51</td>
</tr>
<tr>
<td>4. Managing Personnel</td>
<td>.48</td>
<td>.49</td>
</tr>
<tr>
<td>5. Analysis and Synthesis in Decision-Making</td>
<td>.63</td>
<td>.65</td>
</tr>
<tr>
<td>6. Productivity</td>
<td>.55</td>
<td>.67</td>
</tr>
<tr>
<td>7. Quality of Judgement(^a)</td>
<td>.41</td>
<td>.46</td>
</tr>
</tbody>
</table>

**Note.** \(^a\)Quality of Judgement scores are measured by the Panel 2 key (Conservative method of combining judgements).
Table 19

Bivariate Correlation between Number of High Priority Items Attempted and Overall Management Performance (OMP)

<table>
<thead>
<tr>
<th>Bivariate Correlation with OMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
</tr>
</tbody>
</table>

Number of High Priority Items Attempted | .21* |

Note. Reported $r$ has been corrected for effects of range restriction and criterion unreliability. Non-significant correlations were omitted. *$p < .05$ (one-tailed test).
this new impressionistic measure does not contribute useful predictive information about participants' administrative performance.

Factor Analysis of the TSIB Action Elements

It was desired to construct an empirically rather than logically-derived set of scoring keys for the stylistic aspects of administrative performance. Accordingly, the action elements from a subset of the TSIB items were combined in a factor analysis in order to develop a set of meaningful, independent stylistic dimensions of in-basket performance using factor loadings to assign action elements to performance dimensions (factors).

The same items comprising the reduced-item subsets used in the Quality of Judgement analyses were included in the factor analysis. Specifically, the action elements from largest set of 12 items, which incorporated the same items used in both the 8- and 10-item sets, along with action elements from two additional items, were combined in a factor analysis. In total, 440 action elements from the 14 items were selected as variables. No statistical software package was available which could factor analyze such an unusually large number of variables. Software limitations also precluded an examination of the data from the 440 variables for mean gender differences and for homogeneity of covariance matrices. As a result, separate gender data were pooled and several factorings of smaller sets of selected variables were carried out. In all, three factorings were conducted.

A Series of Factorings

The first factor analysis. In order to reduce the number of variables to permit computer analysis (the least limited software package allowed approximately 275 variables), the set of 440 action elements was first reduced by eliminating those with
low endorsement rates, defined as those action elements endorsed four times or less over the sample of 321 participants. In addition, the action elements from two items (those not part of the 12-item subset) were excluded from first factor analysis, resulting in 269 action elements remaining for inclusion as variables in the factor analysis.

The criterion used to extract the optimal number of factors was primarily rational rather than psychometric. The commonly-used Kaiser-Guttman rule of the number of eigenvalues greater than unity was not appropriate to determine the optimal number of factors because the number of eigenvalues greater than unity was excessive (95, in total). Cattell's (1966) scree test indicated that four factors were likely the correct number. The decision to extract six factors in this first analysis was made with the rationale that fewer factors may not be comprehensive enough to describe the data and a greater number of factors would be difficult to interpret. The six factors were obtained and transformed to an optimal oblique simple structure using the oblimin rotation procedure (Carroll, 1960). Factor loadings from the oblique primary factor-pattern matrix were examined in order to select a subset of variables for inclusion in the second factoring. Variables with loadings of .30 or greater were targeted for retention whereas those variables with loadings less than .30 were excluded in the subsequent factoring.

The second factoring. The variables involved in the second factoring of the data included the targeted variables from the first analysis and the action elements from the remaining two items of the 14 items identified for factor analysis, for a total of 218 action elements/variables. Again, six factors were obtained and transformed to an optimal oblique simple structure by the oblimin rotation procedure. Factor loadings from this second factor pattern matrix were examined to select variables for inclusion in the final factoring of the data. Relatively few large loadings were evident. Accordingly, it was decided to exclude the action elements from the two new items
added in the second factoring, and provide a more complete factoring of the 12-item subset by including those action elements which had been previously excluded because of low endorsement rates.

**The third factoring.** For this final factoring, a total of 268 action elements were retained as variables. Six factors were extracted and transformed to an optimal oblique simple structure by the oblimin rotation procedure. In order to facilitate the interpretation of the factors from the sizable factor-pattern matrix, a grouping of action elements (variables) based on factor-pattern matrix loadings for each factor across the 268 variables within each factor was made: Group "A" action elements with loadings of .26 or greater, and Group "B" action elements with loadings between .175 and .25. The extensive factor-pattern matrix is too large to report here. Instead, the general sense of each factor will be discussed below, with some specific action element content given to make the discussion more concrete.

**Interpretation of the factors**

On the basis of this third oblique factor solution, five interpretable factors emerged, which were interpreted as follows:

**FACTOR I: CONSULTATION AND DISCUSSION.** This factor indicates a tendency to discuss with superiors and subordinates before taking action, to communicate in person with a focus on immediate, short-range concerns and issues, and to require an exchange of views in an attempt to reach a decision. The central trait for persons high on this factor appears to be an orientation to people.

A total of 16 action elements were classified as Group "A" action elements, whereas 26 action elements were considered Group "B" elements. The action elements loading on this factor (listed in Appendix B) show a pattern, across items, of initiating
discussions with various staff rather than making final decisions. Although the content of some action elements (primarily those with B loadings) suggest meetings are not to be held, it is likely that, given the particular item content, action elements calling for discussions and meetings with personnel would be more conclusive, rather than preparatory, in dealing with problem presented in the item. For example, Item 10 consisted of a terse memo from the participant's superior threatening disciplinary action if productivity is not improved. The participant is directed to sign the memo and distribute it to all personnel. It is likely that meeting with staff to discuss the productivity problem is a definitive step toward resolving the issue, and is more action-oriented than holding discussions in the context of issues presented in other items.

FACTOR II: INDEPENDENT DECISION-MAKING. This factor appears concerned with the ability to take final action without guidance from superiors or subordinates and the ability to get things done in an autonomous, decisive way. Persons who score high on this factor would likely not require close supervision or control to make decisions, and would show initiative and drive in identifying and following through with actions.

For this factor, a total of 17 action elements were classified as Group "A" action elements and 17 action elements were categorized into Group "B" elements. The action elements loading on this factor consistently display an action-oriented approach to dealing with the items in the TSIB. There is a tendency to not discuss issues with superiors and to act conclusively in handling problems. For example, the action elements with the largest loadings from Item 4 (reproduced in Figure 4) include finding out who is involved in the coffee break violations, driving by the Coffee Shop, and giving reprimands to those involved.
FACTOR III: COMPANY-BASED DECISION-MAKING. This factor describes one who tends to comply with policies and decisions set out by superiors and who takes action, but likely after consultation with superiors. A tendency to solicit opinions and guidance from higher levels in developing and carrying out managerial duties, a concern with maintaining consistency of individual decisions and actions with the broader mandate from company, and a sensitivity to company priorities characterize those who score high on this factor.

Here, a total of 18 action elements were classified as Group "A" action elements, whereas 19 action elements were considered Group "B" elements. Across items, the action elements with the highest loadings (and thus the most interpretive weight) exhibit a common thread of communication with superiors and a desire to conform to company directives, such as minimizing or prohibiting overtime, despite a backlog of orders.

FACTOR IV: SUPERVISING STAFF. This factor appears related to involvement in scheduling and supervising the activities of staff, the tendency to ask for information and to review and control the allocation of staff. A supervisor scoring high on this factor would likely provide structured, close monitoring of the activities of staff, would delegate and direct work to subordinates, would provide suggestions or proposals to staff to deal with issues or problems, and would be able to disapprove suggestions or plans from staff as necessary.

For this factor, a total of 15 action elements were each classified as Group "A" and Group "B" elements. Across all factors, the highest loadings are seen in this fourth factor. A strong and consistent pattern of actions which involves assigning duties to staff, and meeting with both subordinates and the company resource person (likely to discuss staff conflicts), is evident.
FACTOR V: INTEGRATION AND ANALYSIS. This factor is concerned with the degree of recognition of the inter-relatedness of problems across items or situations, the ability to see broader view and implications of problems in the exercise, and a consideration of the urgency or priority of taking actions or in deciding the order for work.

For this final factor, seven action elements were classified as Group "A" action elements, whereas 33 action elements were considered Group "B" elements. Although this factor yields the lowest number of action elements with A-level loadings, consideration of the B-level loadings suggested a tendency to investigate through peer discussions and more impersonal fact-finding actions (e.g., ask secretary for overtime figures). Of the five final factors, the few large loadings and numerous moderate loadings (33, in total) made this the most difficult factor to interpret.

The sixth factor extracted from the third oblique factor solution did not present a clear, interpretable pattern based on action element content and so was omitted. The intercorrelations among the five primary factors are shown in Table 20.

Assignment of Action Elements to Factors

Using the factor loadings provided by the factor-pattern matrix of the third factor solution, a combination approach of empirically and logically assigning action elements to factors was taken in order to determine factor-scale scores. In general, the empirical results were given primary consideration, in that action elements which were
Table 20

**Primary-Factor Intercorrelations**

<table>
<thead>
<tr>
<th>Factors</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>.01</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>-.02</td>
<td>-.01</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>-.01</td>
<td>-.07</td>
<td>.02</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>.03</td>
<td>-.01</td>
<td>.01</td>
<td>.02</td>
<td>--</td>
</tr>
</tbody>
</table>
factorially simple (i.e., loaded onto one factor) were keyed for that factor. In more factorially complex cases, the typical action element assignment was based on the largest loadings for that element. In more ambiguous cases (factor loadings of similar magnitude), the content of the action element was considered, and the final assignment was based on logically matching the administrative performance involved in the action element to the most closely related (by content) factor. Multiple keying of action elements to factors (i.e., assigning an action element to more than one factor) was minimized in order to retain the distinctive, independent nature of these largely factor-analytically derived dimensions.

It will be recalled that the second factoring included two additional items, not part of the 12-item subset, and that the action elements from these items were subsequently eliminated for the third factoring. However, these action elements were now included in the derivation of the factor-scores in order to increase the number of action elements used and increase the variability of the new dimensions. To this end, a combination process of action element assignment identical to that just described was used, except that factor loadings from the primary factor-pattern matrix from the second factoring were studied to assign the additional 69 action elements to factors in the computation of factor-based scale scores. Finally, to increase the base of measurable behaviours for Factor V (Integration and Analysis), the conceptually-salient variable of the Number of Priority Items Attempted was added in the calculation of Factor V scores. Like the logically-derived stylistic dimension scores, final scale scores were derived by summing the number of actions elements with positive loadings coded for each factor and subtracting those with negative loadings that were coded for each factor.

The scale score intercorrelations are provided in Table 21. The results of the correlations between these scale scores and the OMP, presented in the upper portion of
Table 21

Factor-Based Scale Score Intercorrelations

<table>
<thead>
<tr>
<th>Factors</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consultation and Discussion</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent Decision-Making</td>
<td>-.05</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Company-Based Decision-Making</td>
<td>.00</td>
<td>-.07</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervising Staff</td>
<td>.21*</td>
<td>.07</td>
<td>.10</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Integration and Analysis</td>
<td>.10</td>
<td>.01</td>
<td>.00</td>
<td>.13</td>
<td>--</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consultation and Discussion</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent Decision-Making</td>
<td>.05</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Company-Based Decision-Making</td>
<td>.09</td>
<td>-.08</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervising Staff</td>
<td>.19*</td>
<td>.02</td>
<td>.00</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Integration and Analysis</td>
<td>.00</td>
<td>.00</td>
<td>.10</td>
<td>.29</td>
<td>--</td>
</tr>
</tbody>
</table>

Note. *p < .05 (one-tailed test).
Table 22

Validity Coefficients between Overall Management Performance (OMP) and TSIB Factor-Based Scale Scores and Reliability Estimates

<table>
<thead>
<tr>
<th>Factor</th>
<th>Bivariate Correlation with OMP</th>
<th>Alpha Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Consultation and Discussion</td>
<td>.18*</td>
<td></td>
</tr>
<tr>
<td>Independent Decision-Making</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Company-Based Decision-Making</td>
<td>.18*</td>
<td></td>
</tr>
<tr>
<td>Supervising Staff</td>
<td>.18*</td>
<td></td>
</tr>
<tr>
<td>Integration and Analysis</td>
<td>.24**</td>
<td></td>
</tr>
</tbody>
</table>

Note. Reported $r$'s have been corrected for the effects of range restriction and criterion unreliability. All non-significant correlations were omitted. 
*p < .05 (one-tailed). **p < .01 (one-tailed).
Table 22, were not promising. Although 4 of 10 possible correlations showed a significant relationship with on-the-job managerial performance, the coefficients were very modest. None was significant for both genders. In addition, the reliability estimates (alpha coefficients) for the scale scores, presented in the lower portion of Table 22, showed strong evidence of unreliability.
DISCUSSION

In general, mixed support was found for the main hypotheses proposed in the present study. In this section, because of the scope and number of hypotheses involved, they will be briefly reviewed preceding the evaluation of the results, in order to more fully interpret and consider the implications of the findings.

Reliability and Criterion-related Validities of Selected Dimensions of the TSIB

Cross-Validation of the Quality of Judgement Scores

Predictive validities of scores based on the Panel, Empirical and Merged keys. A central focus of the present study was the degree to which the three Quality of Judgement keys (Panel 2--Conservative, Empirical, and Merged), as applied to the three reduced-item subsets (8-, 10-, and 12-item) would yield similar criterion-related validity results when administered in a second, cross-validation sample (Company 2). It was hypothesized that the Merged and, to a lesser degree, the Empirical key used to measure the Quality of Judgement dimension would result in the largest validities for the three scoring keys and that there would be some shrinkage in the cross-validated coefficients based on the Empirical and Merged keys.

The findings, presented in Table 9, showed that these expected results were only partially realized. The Merged key, across both genders and for each of the three reduced-item subsets, did yield higher validity coefficients than the Empirical key, with coefficients ranging from .18 to .23 (corrected). It should be noted, however, that only two of six Empirical key correlations were significant (8-item set and 10-item set, \( r = .19 \)). Even though some shrinkage in the cross-validated Quality of Judgement scores was hypothesized, such a marked decrease using the Empirical key was unexpected. Serious limitations in the empirically-based key were revealed in that four
of the six reduced-item sets measured using the Empirical key were found to be unrelated to on-the-job measures of managerial performance. Furthermore, the two significant correlations were observed for females only, therefore no predictive information about males' managerial performance is provided by the Empirical key. It should also be noted that, although all six possible cross-validation Merged key correlations were significant, the average magnitude of the correlations from the second dataset was substantially lower. For example, the mean Merged key validity coefficient (across both genders and the three reduced-item subsets) generated using the Company 1 dataset was .61, whereas the mean validity coefficient based on the Company 2 dataset was .21.

These reductions in the validity of the Quality of Judgement scores as measured by both the Empirical and Merged keys are likely seen because the original validity coefficients reported in Preliminary Study 2 were spuriously inflated from the effects of capitalization on chance, which produces statistically significant correlations simply on the basis of chance, or Type I error. The considerable number of correlations required in the determination of the Quality of Judgement values for these two keys (228 correlations for the 8-item subset, 291 for the 10-item subset, and 354 for the 12-item subset) presumably increased the likelihood that chance factors could operate. Another potential factor contributing to the inflation of the original validity coefficients was the fact that the reduced-item subsets were chosen largely on the basis of maximal item validities (rather than reliabilities) for the original sample. In addition, it may have been that, in the correlational analyses used to develop these scoring keys, the inclusion of correlations based on action elements with low endorsement rates (e.g., those endorsed by 5 of 321 participants) did not yield stable correlations. Regardless of the reasons for the over-estimates of the original validity coefficients, it is clear that the second, more accurate assessment of the true validity of the Quality of Judgement
dimension as measured by the Empirical and Merged keys shows much less promising findings.

Contrary to expectations, the Panel 2 (Conservative) key yielded the highest validity coefficients of the findings from the three main approaches used in Quality of Judgement scoring key development and reported in Table 9. Unlike the cross-validated coefficients from the Empirical and Merged keys, no shrinkage was seen in the Panel 2 (Conservative) validity coefficients; in fact, the average magnitude of the six correlations from the second dataset was marginally higher. Specifically, the mean Panel 2 (Conservative) key validity coefficient (across both genders and the three reduced-item subsets) generated using the Company 1 dataset was .255, whereas the mean validity coefficient based on the Company 2 dataset was .273. It was expected that the Panel 2--Conservative key would result in at least equal, if not slightly higher, validity coefficients across the two datasets. Whereas the original analysis represented an inter-company analysis (the origin of the panel and the dataset did not match; Panel 2 applied to Company 1 dataset), the cross-validation represented an intra-company analysis (the origin of the panel and the dataset match; Panel 2 applied to Company 2 dataset). Higher validity coefficients from intra-company analyses were expected because confounding effects due to differences in corporate culture or geographical location were presumed to be minimized, if not eliminated.

A final point in the consideration of the Panel approach is that it may have been responsible, at least in part, for the higher coefficients seen with the Merged key compared to the Empirical key. It will be recalled that although the derivation of the Empirical key was solely quantitative, in contrast the derivation of the Merged key incorporated both the Empirical and the qualitative Panel judgements. It is likely that the inclusion of the more predictive Panel judgements in the Merged key increased the observed cross-validated validity coefficients. In sum, it is evident that subjective,
human judgements, rather than objective, empirical calculations provided more enduring, accurate estimates of the predictive validity of the Quality of Judgement scores.

In retrospect, the expectation that the Panel 2 (Conservative) key would yield the lowest validity coefficients from the three main approaches used in Quality of Judgement scoring key development may not have been called for, particularly in light of the findings of Stern, Stein, and Bloom (1956). Stern et al. developed an important classification of three approaches used in personality assessment programs, based on the extent to which an explicit personality model is used in predicting human behaviour. Interestingly, the analytic, empirical, and synthetic approaches identified by Stern et al. closely correspond to the three approaches of the Panel, Empirical, and Merged Quality of Judgement scoring key development used in the present study. Comparable to the Panel approach, the analytic approach to personality assessment is complex. At its simplest level, however, it involves gathering and combining individual ratings of the role requirements made by significant others. A further step in the analytic approach involves a diagnostic council, whose purpose is to assess, by group consensus, the congruence of a candidate's personality with the hypothetical target model. According to Wiggins (1973), the most outstanding advantage of the analytic approach is its generality of predictive accuracy across samples of both individuals and situations.

The empirical approach identified by Stern, Stein, and Bloom (1956) is based on examination of the degree of correlation between instruments selected for personality prediction and objective standards of performance. Item analysis following a contrasted groups application to high- and low-performing subjects is used to select items able to reflect group differences. In this approach, as in the Empirical Quality of Judgement scoring key approach, cross-validation is required to avoid capitalization on chance influences from the original contrasted groups. The view that "the degree of success of
empirical selection is considered to be less than optimal in terms of such considerations as hit rates, costs of testing, and selection ratios” (Wiggins, 1973, p. 469) is supported by the findings of the present study. In addition, Wiggins noted that the generality of the empirical approach is much less than that of the analytic and synthetic approaches.

The synthetic approach, like the Merged key, involves a more global or armchair appraisal of the situation, combining assessors' ratings in a less rigorous way than the empirical approach. It should be recognized, however, that the Merged key approach is less intuitive and follows more clearly established decision rules than the process described by Stern, Stein, and Bloom (1956). Generally, the synthetic approach is considered less accurate than the empirical approach in predicting personality, and it is considered more efficient than the comprehensive analytic approach.

In sum, given the results of Stern, Stein, and Bloom (1956), the finding that, of the three main Quality of Judgement scoring key approaches, the Panel 2 (Conservative) key yielded the highest validity coefficients and is most generalizable is considerably less contrary than originally expected. Despite obvious differences in domain assessment, there is an interesting convergence to the pattern of findings across the two studies.

Additional analyses based on the Panel scoring key. The administration of the TSIB in a new, related setting allowed a set of more finely-grained analyses of the Panel scoring keys used in determination the validity of the Quality of Judgement scores. It will be recalled that the involvement of Company 2 permitted the construction of a new, expanded panel. To increase the accuracy and stability of the appropriateness ratings, 21 panel members were used for Panel 2, compared to 11 panel members comprising the first panel (from Company 1). In addition to more panel members from a different company, Panel 2 also differed from Panel 1 by the method of combining panel judgements (Panel 1 used the liberal method while Panel 2 used the conservative
method). These multiple differences in the two panels, then, would not allow the isolation of the factor(s) responsible for differing validity coefficients resulting when the two panels were used in the derivation of the Quality of Judgement scoring key. More precise comparisons of results required the separation of these three factors: the method of combining panel judgements (liberal or conservative), origins of the panel ratings (Panel 1 or Panel 2), and the dataset to which the particular panel key is to be applied.

Accordingly, the six additional Panel analyses previously outlined in the Rationale and Hypotheses for the Present Study were conducted. (It will be recalled that Study 1 involved the Liberal Method using Panel 1 applied to the Company 1 dataset and Study 2 involved the Conservative Method using Panel 2 applied to the Company 1 dataset.) To review, the new analyses involved:

1. Conservative combination method using Panel 1 applied to Company 1 dataset.

2. Conservative combination method using Panel 1 applied to Company 2 dataset.


4. Liberal combination method using Panel 1 applied to Company 2 dataset.

5. Liberal combination method using Panel 2 applied to Company 1 dataset.


It was hypothesized that, because of the greater consensus required among the panel members and the use of a 5-point, rather than 3-point, scale, the Conservative method of combining judgements would result in greater validity coefficients than the
Liberal method. In addition, it was expected that, in general, Panel 2 ratings would yield higher validities than Panel 1 ratings because of the use of expanded list of action elements and the larger panel size used for the action element ratings. Lastly, it was believed that "intra-company" analyses, wherein the origin of the panel and the dataset match, (e.g., Panel 1 applied to Company 1 dataset; Panel 2 applied to Company 2 dataset) would produce greater validity coefficients than "inter-company" analyses (Panel 1 dataset applied to Company 2 dataset). This expected pattern was based on the assumption that intra-company analyses would tend to eliminate possible confounding effects arising from differences in corporate culture or geographical location. Table 23 presents a summary of the observed validity coefficients reported in the previous section.

In sum, the effects of three factors on the validities of the Quality of Judgement scores were examined: the method of combining panel judgements, the origin of the Panel, and the nature of the analysis (whether intra- or inter-company). The analyses conducted in the present study were restricted to a consideration of the main effects of these factors; no interactive effects were investigated. Table 24 presents the means of the observed validity coefficients, aggregated over each of the three factors and for each of the three item subsets. The means, therefore, were determined by a broad aggregation of values shown in Table 23. For example, in the calculation of the Liberal method means, the categories of gender, origin of the Panel, and intra- versus inter-company analysis were collapsed. Similarly, in the calculation of the Panel 2 means, the categories of gender, method of combining panel judgements, and the intra- versus inter-company analyses were collapsed.
Table 23
Summary of Validity Coefficients between Overall Management Performance and Quality of Judgement Scores Obtained Using Scoring Keys based on Different Methods of Combining Panel Judgements (Liberal or Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), and Whether Panel Origin and Dataset Match (Intra- or Inter-company analysis)

<table>
<thead>
<tr>
<th>8-item set</th>
<th>Liberal Method</th>
<th>Conservative Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Panel 1</td>
<td>Panel 2</td>
</tr>
<tr>
<td>Intra-company</td>
<td>M  .37</td>
<td>F  .29</td>
</tr>
<tr>
<td>Inter-company</td>
<td>M  .26</td>
<td>F  .35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10-item set</th>
<th>Liberal Method</th>
<th>Conservative Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Panel 1</td>
<td>Panel 2</td>
</tr>
<tr>
<td>Intra-company</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Inter-company</td>
<td>M  .23</td>
<td>F  .35</td>
</tr>
</tbody>
</table>
Table 23 (cont.)
Summary of Validity Coefficients between Overall Management Performance and Quality of Judgement Scores Obtained Using Scoring Keys based on Different Methods of Combining Panel Judgements (Liberal or Conservative), Different Origins of Panel Ratings (Panel 1 or Panel 2), and Whether Panel Origin and Dataset Match (Intra- or Inter-company analysis)

<table>
<thead>
<tr>
<th>12-item set</th>
<th>Liberal Method</th>
<th>Conservative Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Panel 1</td>
<td>Panel 2</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Intra-company</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Inter-company</td>
<td>.18</td>
<td>.36</td>
</tr>
</tbody>
</table>

Note. Reported r's have been corrected for the effects of range restriction and criterion unreliability. Correlations for the 10- and 12-item sets, scored by the Liberal Method, Panel 1 key, applied to the Company 1 dataset are not available. Value in parentheses is a non-significant correlation.
Table 24

**Mean Validity Coefficients Determined by Aggregation Across Method of Combining Panel Judgements, Origin of Panel, and Nature of Analysis (Intra- or Inter-company)**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8-item</td>
</tr>
<tr>
<td><strong>Method of Combining Panel Judgements</strong></td>
<td></td>
</tr>
<tr>
<td>Liberal</td>
<td>.31</td>
</tr>
<tr>
<td>Conservative</td>
<td>.27</td>
</tr>
<tr>
<td><strong>Origin of Panel</strong></td>
<td></td>
</tr>
<tr>
<td>Panel 1</td>
<td>.29</td>
</tr>
<tr>
<td>Panel 2</td>
<td>.29</td>
</tr>
<tr>
<td><strong>Nature of Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Intra-company</td>
<td>.29</td>
</tr>
<tr>
<td>Inter-company</td>
<td>.29</td>
</tr>
</tbody>
</table>

**Note.** Means were determined by Fisher transformation of the original bivariate correlations.
As shown in Table 24, no support was found for the hypothesis of higher validity coefficients derived from the Conservative, rather than Liberal, method of combining panel judgements. Across the three item-subsets, the mean validity coefficient for the Conservative method was .26, compared to .28 for the Liberal method. It is evident that requiring greater consensus among panel members and the use of a 5-point, rather than 3-point scale, in the assignment of Quality of Judgement values did not result in improved predictive validity of the Quality of Judgement scores. This finding suggests that, because of simpler, more efficient computations, the Liberal method of combining panel judgements is the method of choice.

Limited support was found for the hypothesis that Panel 2 ratings would yield higher validity coefficients than Panel 1 ratings. The coefficients were higher for all three possible comparisons for males whereas for females, none of the three possible comparisons was higher (the pattern was fully opposite to that predicted). For the males, then, it is apparent that using the expanded list of action elements and the larger panel size resulted in greater predictive accuracy (8-item, .30 versus .28; 10-item, .24 versus .22; 12-item, .25 versus .21). The question remains whether, given the marginal increase in validity coefficients for males using Panel 2 and the decrease in validities for females, the use of the expanded list of action elements and a panel size nearly double that of the original panel is warranted. Moreover, using the validity coefficients from Table 24, which have been aggregated across genders, the means of the validity coefficients across the three item-subsets are equal for the two Panels. Practical considerations point to the use of Panel 1 as the more favored of the two panels. That is, using a smaller (rather than larger) panel and a less comprehensive list of action elements will yield similar validity coefficients with the added advantage of reduced cost and time demands associated with the smaller panel. It should be noted that, although Panel 1 is referred to as the smaller of the two panels, when compared to panel sizes reported in the literature
(Hakstian et. al, 1986) and those employed by ETS, Panel 1, with 11 members, is relatively large.

Limited support was also found for the hypothesis that intra-company analyses would yield greater Quality of Judgement validity coefficients than inter-company analyses. Like the previous analyses comparing Panel 1 to Panel 2 ratings, a differential pattern across genders was seen (Table 23). The coefficients were higher for all three possible intra-company comparisons for males, whereas for females, none of the three possible intra-company comparisons was higher (again, the pattern was fully opposite to that predicted). When considered across genders (Table 24), matching the origin of the panel to the dataset to which the Quality of Judgement key will be applied does not appear to be a requirement for satisfactory validity. Specifically, across the three item-subsets, the mean validity coefficient for the intra-company analyses was .27 whereas the mean for the inter-company analyses was .28. It can be seen that, overall, no significant reduction in validity results if the panel and dataset do not correspond, and so it is unlikely that differences due to corporate culture or geographical location have appreciable confounding effects on the Quality of Judgement key. An important consequence of this finding is the implied versatility of panel-derived keys. It appears that panel keys derived in the original, target company can also be applied in new, related settings without significant reductions in validity.

**Additional Performance Dimensions in the TSIB**

**Understanding of Situation Questionnaire.** In the present study, a new eighth dimension (performance rather than stylistic) was added in order to measure the participant's understanding of the issues, problems and broader implications of the scenario presented in the TSIB. It was hypothesized that more information about a participant's rationale or thinking would constitute useful predictive information about
that person's administrative potential. Because of the fairly low internal consistency results in Study 2, it was postulated that the Quality of Judgement dimension may be made up of multidimensional, independent aspects of appropriateness of actions. A 33-item multiple choice Understanding of Situation Questionnaire was designed and administered upon completion of the in-basket exercise.

The development of the scoring key for the Understanding of Situation Questionnaire was based on an analysis of the responses from the second 21-member panel involving the application of a 0, 1, 2 scoring system assigned to each of the five options per question. It will be recalled that the particular value assigned to each option depended on the level of endorsement from the expert panel. Across the five response options provided for an item, the panel members' pattern of option selection was examined and an empirical, followed by a logical approach, was used to assign scoring weights. The empirical approach was based on the frequencies of option selections within an item. If, for example, option "b" were selected most frequently by panel members, it was initially keyed "2" whereas if option "d" were chosen by very few members (or none), it was initially keyed "0". Options receiving moderate endorsement by the panel (i.e., options neither maximally nor minimally selected) were initially keyed "1". Following this empirically-based assignment of scoring weights, final determination of the scoring key was made by a logical or rational consideration of the relative differences between option selection rates. If, for example, the two most frequently selected options within an item differed by one or perhaps two panel members' selection, both options were keyed "2". Similarly, if two moderately-endorsed options showed a minimal difference in their endorsement rates, they both were keyed "1".

As outlined earlier in the Rationale and Hypotheses for the Present Study, item-analyses using item-reliability and item-validity indices were conducted in order to select a subset of items from the questionnaire with maximal reliability and validity. Table 11
presented the internal consistency and criterion-related validity findings based on the original, full 32-item set (one item was omitted on the advice of the panel). Table 12 presented the internal consistency results and validity coefficients for the most promising subset of items (20, in total). The reliability results indicate low levels of internal consistency for both the full- and reduced-item questionnaire (alpha coefficients for the full-item set were .24 and .38 for males and females, respectively, compared to .23 and .27 for the 20-item subset). However, given the nature of this data, it may be unrealistic to expect results indicating greater homogeneity. The questionnaire is designed to solicit information regarding participants' attitudes, strategies and rationales behind the numerous and varied actions they chose. The data, therefore, is more cognitive and item or issue-specific than is typically true of more general, task-oriented in-basket data.

Based on the criterion-related validity findings from the 20-item questionnaire, strong support was found for the hypothesis that additional, useful predictive information which augments in-basket assessment could be provided by an analysis of participants' particular approaches in the exercise. The criterion-related findings suggest that the 20-item Understanding of Situation Questionnaire not only yields higher, but also more equal, validity coefficients across genders compared to the original, full-item instrument (32-item, .19 and .29 for males and females, respectively; 20-item, .35 and .34 for males and females, respectively). An important practical advantage of the questionnaire format compared to the standard ETS Reasons for Action form or a post-exercise interview, designed to solicit the same information, is afforded with this multiple-choice scoring format. Specifically, less administration and scoring-time is required and improved reliability (inter-rater) follows from using a quantitative, objectively-applied scoring key.

**Productivity.** Findings from the preliminary studies revealed that the Productivity dimension did not contribute significantly to the prediction of in-basket performance for men or women. It will be recalled that Productivity was operationalized
as the total number of actions taken across the exercise (including Unusual actions), as well as the number of items attempted. Consequently, in the present study, several different ways of operationalizing Productivity to increase the validity of this dimension were examined.

New methods of defining this dimension involved the aggregation of smaller, more concrete units of output. Seven separate units of productivity were identified as individual productivity components: (a) the number of items attempted across 21 items, referred to as Item Total; (b) the number of letters or memos written across 21 items, referred to as # Letters/memos; (c) a score reflecting the number of words written per memo or letter across the 21 items, referred to as Total # words written; (d) number of actions scheduled for a definite time, referred to as Definite actions; (e) number of entries made on the calendar; (f) whether a "things to do" list was completed, and (g) whether a "summary" list of the items was completed. In addition, more standard indices of output were calculated for each item-subset, composed of the total number of action elements and total number of Unusual actions taken over the 8-, 10-, and 12-item subsets. These were referred to as Total action (8), Total action (10), and Total action (12), respectively.

Criterion-related validity results for the individual units of output, presented in Table 13, show a different pattern across genders. For males, the # Letters/memos written (across 21 items) and the Total # words written (also across 21 items) were significantly related to on-the-job managerial performance (r = .24 for both measures). For females, all individual units of output were non-significant, except whether a summary list was completed by the participant (r = .20).

Productivity linear composites were then derived to determine the optimal Productivity operationalization. The most promising individual units of output were selected and four linear composites were determined: (a) Item Total, # Letters/memos,
Total action (10); (b) Item Total, Total # words written, Total action (10); (c) Item Total, Definite actions, Total action (10); and (d) Item Total (across 10 items), # Letters/memos, Total action (10). As Table 13 showed, the fourth linear composite, with significant validity coefficients of .25 for both genders, proved to be the optimal operationalization of Productivity, although the first linear composite, with significant validity coefficients of .24 for both genders, was a close second. The former linear composite, however, has the added advantage of being the more easily scored of the two composites, because the Item Total is calculated across 10, rather than 21, items. For comparison, a more conventional approach to the scoring of Productivity was examined in the Total action measures for the 8-, 10-, and 12-item subsets. The Productivity totals for the 10- and 12-item sets yielded the highest coefficients ($r = .18$ for males and $r = .25$ for females, for both item-sets).

The results of these analyses suggest that measuring Productivity using the composite based on the Item Total (across 10 items), # Letters/memos across the entire exercise, and the Total action (10) count provides the most predictive information about administrative potential (mean validity coefficient is .25). However, it should be recognized that the mean validity coefficient, across genders, for the Total action (10) measure is only marginally lower ($r = .22$) than this linear composite. Practical considerations point to the selection of this second, slightly less predictive operationalization of Productivity as the optimal method of measuring in-basket output. Unlike the linear composite, the scoring of the standard indices of output require no additional steps or calculations other than those which follow directly from scoring the in-basket itself. That is, in the scoring of the TSIB, the action elements (both Unusual ones and those listed in the Scoring Manual) are identified. The total number of action elements and the total number of Unusual actions taken (whether using the 8-, 10-, and 12-item subset), already determined in the scoring of the TSIB, are then summed,
making Productivity assessment a straightforward computation. In contrast, to score the linear composite, an additional component of the number of letters and memos written across the entire exercise must be counted. This component, although not complex, is fairly time-consuming to determine because all written responses, not simply those relevant to the subset of items, must be examined. Thus, with such practical advantages, it appears that the Total action measure is the most useful method of measuring in-basket productivity; the ultimate selection of the Total action (8, 10, or 12) rests on the selection of the optimal reduced-item subset.

**TSIB Stylistic Dimensions**

The efficacy of logically assigning the stylistic dimensions was re-examined in the second dataset using the 8-, 10-, and 12-item subsets. The total scores for each stylistic dimension were correlated with the OMP for each of the three item-subsets and were displayed in Table 14. As reported in the Results section, a distinct pattern across genders, but consistent within each gender, was evident. It will be recalled that, across the item-subsets, Interpersonal Relations and Leadership in a Supervisory Role were the only significant stylistic dimensions for men, whereas Planning and Organizing Work, Managing Personnel, and Analysis and Synthesis in Decision-Making provided useful predictive information for women.

These results using the reduced-item subsets can be compared, generally, with those from Study 1 wherein the full 21-item set was administered in Company 1 (results displayed in Table 5). There, greater consistency across genders in the results was seen with two stylistic dimensions, Interpersonal Relations and Managing Personnel, significant across both genders. In the present study, it appears that only a small set of stylistic dimensions, unique to each gender, were related to measures of on-the-job performance. It is not clear why such a differential pattern across genders should exist.
It is interesting to note that not one of the five stylistic dimensions is entirely nonsignificant; each dimension contributes some useful predictive information about managerial performance, albeit for only one gender.

Tables 15, 16, and 17 present the intercorrelations for the TSIB stylistic dimensions for the three reduced-item subsets. Consistent with the findings from the intercorrelations of the 8-item subset reported in Study 2 (Table 7), a fairly high level of intercorrelations for the reduced-item subsets was seen, suggesting a lack of meaningful discriminant validity of these aspects of administrative performance. This dimensional dependency most likely results from multiple keying of action elements across dimensions. Earlier studies using such multiple keying of logically-assigned dimensions also found evidence of dimensional dependency (Brannick et al. 1989; Hakstian & Harlos, 1992).

Reliability of the TSIB dimensions

The split-half reliability estimates of the seven dimensions, presented in Table 18, fall in the middle to upper part of the range of split-half reliabilities reported by Schippman et al. (1990) and summarized in the Table 1 of the Literature Review. It was noted in the Results section that a pattern of increasing reliability estimates with larger item-subsets was seen, supporting the speculation by Hakstian et al. (1992) that the phenomenon of test length is responsible for the lower split-half reliability of individual TSIB dimension scores for the 8-item subset compared to higher reliability estimates for the full 21-item set. Thus, as expected, the more items scored, the greater the split-half reliability. Similarly, the reported alpha coefficients, presented in Table 18, also show a pattern of increasing magnitude as the number of items increases. For example, for males the mean alpha coefficients across the seven dimensions and the three item subsets
were .30 (8-item), .39 (10-item) and .47 (12-item). For females, the mean alpha coefficients were .38 (8-item), .45 (10-item) and .54 (12-item).

It is apparent that, regardless of the reliability estimation method used and the number of items scored, the TSIB dimension scores, like in-basket dimension scores reported in the literature, are not highly internally consistent. The Literature Review observed that there was great variability in the published split-half reliability coefficients, making it difficult to conclude, with confidence, whether or not the split-half reliabilities are satisfactory. Less variability was seen in the alpha coefficients reported in the literature. The mean alpha coefficient, across in-basket studies, was .51. In the present study, only the 12-item subset reached a comparable degree of homogeneity; the 8-item subset, with a mean alpha coefficient across genders of .34, was substantially lower. Although more promising, published internal consistency findings do not clearly indicate solid levels of reliability using estimates of equivalence and homogeneity. As observed in the Literature Review, the most appropriate measure of the reliability of the in-basket would likely be a measure of stability, obtained by the test-retest method. Yet, because of logistical difficulties and high costs associated with multiple test administrations in industry, assessment of test-retest reliability of the in-basket is generally not viable. Given these limitations, it may be unrealistic to expect stronger evidence of in-basket reliability. If we consider some central features of wideband, high-fidelity in-baskets, such as the breadth and complexity of items, cases of low item variances (due, in large part, to low endorsement rates of courses of action), and the numerous unquantifiable influences that affect the participant's perception of the items, the present findings may appear less negative.
Additional New Measures

Scorer's Impression of Involvement. It will be recalled that two new measures, the Number of Priority Items Attempted and the Scorer's Impression of Involvement, were introduced in the TSIB to investigate their predicted efficacy in improving in-basket assessment. Although previous findings of impressionistic measures of in-basket performance have shown mixed evidence of criterion-related validity (Cross, 1969; Hemphill, 1962; Meyer, 1970), the present attempt to develop a measure which was both subjective and predictive was expected to be favorable because of the clear conceptual link to Lopez's (1966) key notion of ego-involvement in in-basket assessment. Contrary to prediction, the subjective Scorer's Impression of Involvement was not significantly related to on-the-job managerial performance. Consequently, no useful predictive information about participants' administrative performance was contributed by this measure. Previously reported positive findings (albeit modestly so) have typically involved impressionistic measures in the form of adjective-pairs which describe subjective aspects of in-basket performance. It is not clear what unique information these measures contribute for those in-baskets also employing the more standard and predictive ETS stylistic dimension measurement. Therefore, it appears that inclusion of impressionistic measures in traditional in-basket assessment is unwarranted.

Number of Priority Items attempted. Table 19 presented the results of the correlation between the Number of High Priority Items attempted. Mixed support was found for the hypothesis that this measure would be significantly and positively related to managerial performance in that this prediction proved true for males only ($r = .21$). It may be that, with a total of six items identified as high priority, and the fact that most participants complete most of the exercise in the time provided, the variability for the range of possible scores was insufficient. Descriptive statistics for this measure support
this view for, across the sample of 321 participants, the mean score was 5.5, with a standard deviation of .9. It appears, then, that in its present form, a measure of the number of urgent or high priority items completed is too weak to warrant inclusion in the TSIB.

Selection of the Optimal Item Subset:
Integrating Empirical Results and Practical Issues

The selection of the optimal item subset (8-, 10-, or 12-item) based on these reliability findings point to the largest item set with its maximal reliability. However, additional considerations of validity and scoring-time must also be taken into account in the final selection of the optimal item subset. The optimal choice of item subsets in relation to the validity of the Quality of Judgement scores clearly depends on the method used to score this crucial dimension. Following the conclusions reached earlier that, in terms of efficiency and accuracy, the Quality of Judgement dimension is best measured by the Liberal method of combining panel judgements using Panel 1 ratings, this scoring approach will be adopted for the current consideration of the optimal item subset.

It was also seen earlier that no significant reductions in validity resulted when the origin of the panel and the dataset did not match (inter-company application). For this reason, the particular Quality of Judgement validity coefficients selected are based on the Liberal method of combining Panel 1 judgements applied to the Company 2 dataset (no results are available for the 10- and 12-item set of the Liberal method of Panel 1 judgement ratings applied to Company 1). When averaged across genders, these mean validity coefficients (taken from Table 24) were .31 (8-item subset), .27 (10-item subset), and .27 (12-item subset). These findings suggest, then, that the optimal reduced-item subset, in the context of Quality of Judgement validity, is the 8-item set.
As Table 14 shows, the validity coefficients between the TSIB stylistic dimensions and the management criterion are very stable across the item subsets. It appears that no appreciable advantage to the prediction of administrative ability is realized with the selection of the 10- or 12-item subsets.

Clearly, practical considerations affirm the selection of the smallest subset of items. As reported in the findings from Study 2, scoring time using the 8-item key was 23 minutes, on average, and the time required to train scorers was three days. Although it is unlikely that training for scoring the 10- or 12-item set would be substantially longer, estimates for the time required to score the longer subsets are 30 and 40 minutes, respectively. For small-scale applications, the addition of seven or even 17 minutes required for scoring each exercise may not have serious consequences. However, the feasibility of wide-scale applications of the exercise with large numbers of candidates would likely be limited, particularly if the 12-item set were scored.

Factor Analysis of the TSIB Action Elements

It was postulated that, with the high intercorrelations among the stylistic dimension scores seen in Study 2, a factor analysis of the action elements from a subset of the TSIB items would yield a set of meaningful, more independent stylistic dimensions of in-basket performance. In addition, it was expected that these factor-analytically derived in-basket dimensions would help resolve the conflict seen in the literature of the complexity of in-basket performance (whether uni- versus multi-dimensional, in nature), by considering the number of significant factors and the degree to which they are consistent with factors seen in past studies.

It will be recalled that five interpretable factors emerged from this analysis: Consultation and Discussion, Independent Decision-Making, Company-Based Decision-
Making, Supervising Staff, and Integration and Analysis. The primary-factor intercorrelation results, shown in Table 20, provided evidence of minimal interrelations among the factors (the largest correlation was -.07). The intercorrelations of the scale scores, seen in Table 21, suggests that the independent nature of the dimensions was retained. Criterion-related validity results were less promising than expected, with very modest coefficients significant in 4 of 10 possible correlations. As Table 22 indicates, the strongest relationship with on-the-job managerial performance was seen by the Integration and Analysis factor for women (r = .24, corrected; p < .01). The remaining three coefficients showed weaker evidence of a relationship to managerial performance (r = .18, corrected; p < .05), none of which was consistently significant across genders. No consistent evidence of the reliability of the scale scores was seen, with only the Integration and Analysis factor showing a modest level of internal consistency.

Several reasons likely contributed to the lack of solid criterion-related validity and reliability seen in these factor-analytically derived in-basket dimensions. It will be recalled that, in previous factor-analytic studies of in-basket performance, the unit of analysis was the in-basket style category, thus providing molar-level data. The present study, in contrast, employed action element endorsements as the unit of analysis, providing very molecular-level data. A likely consequence of the vast numbers of action elements used (recall that 268 were used as variables in the third factoring) was low within-factor variance due, in large part, to low endorsement rates of many of the action elements. In addition, it is likely that correlations based on action elements which have been endorsed by only a small number of participants (e.g., 5 of 321) were not stable. The very molecular-level approach used in the present factor analysis allowed the measurement of a greater scope or range of in-basket performance compared to more molar-level measurement. This greater scope, however, resulted in reduced frequencies for each molecular unit of measurement, the action element endorsement; a longer list of
action elements yielded greater disbursement of endorsements with fewer endorsements per action element. Consequently, it may be that the use of more molecular-level data in the present study limited the robustness of possible empirical relationships with criterion measures.

Present findings provided some evidence of consistency with factors previously identified in past studies. The Literature Review observed that factor-analytic studies have suggested some common dimensions of administrative performance: Discussing Problems with Others, Complying with Suggestions, Preparing for Action by Gathering More Information, and Directing Others (Thornton & Byham, 1982). In the present study, conceptual overlap is apparent between Consultation and Discussion and Discussing Problems with Others, and between Company-Based Decision-Making and Complying with Suggestions. Moreover, it appears that Preparing for Action by Gathering More Information and Directing Others are two smaller stylistic components of the present factor Supervising Staff.

In the Literature Review, the interpretation of factor consistency across in-basket studies was seen as controversial. Whether the further evidence of factor consistency provided by the present study indicates that the underlying characteristics measured by the in-basket exercise are based on a single, generalized trait, as suggested by Kesselman et al. (1982) and Lopez (1966), remains unclear. If in-basket performance were truly unidimensional, the consistent pattern of three to five distinct, independent factors emerging from solitary studies (regardless of the unit of analysis used--style categories or action elements) would not be expected. Instead, the emergence of one, or perhaps two, unique factors would more likely be seen across past studies. More support is seen for the contention by other researchers that the factor consistency observed in the literature is the result of an ETS-dominated measurement system (Schippmann et al., 1990). This view holds that a similar measurement system may result in greater similarity of factor
definitions, increasing the likelihood for factor consistency across studies. With the strong influence from ETS in the development of the TSIB, such an explanation for the factor consistency observed here seems readily acceptable.

A Note Regarding A Linear Combination of TSIB Dimensions

In factor analysis, optimal combinations of variables are derived from clusters resulting from the internal structure of the data. Other methods to derive optimal linear combinations of variables typically involve correlation of the variables with external criteria, such as job performance. Additional research (Hakstian & Harlos, 1992) into the construction of a simple linear composite of TSIB dimensions has been conducted and will be briefly acknowledged here.

Using the Company 1 dataset from Preliminary Study 1, Hakstian and Harlos (1992) reported that, based on criterion correlations, a linear combination of the TSIB dimensions was constructed, with the Quality of Judgement and Managing Personnel dimensions given weights of two and the other five dimensions given unit weights. This new measure, called Overall In-Basket Performance, yielded significant validities of .30 (males) and .34 (females) when correlated against the Overall Management Performance criterion. When applied in the Company 2 dataset, the cross-validities for these two variables were .28 (males) and .30 (females). The development of the Understanding of Situation dimension allowed it to be added to the linear composite, with a weight of two to reflect its stature as a performance, rather than stylistic, dimension. Consequently, the validities for this global composite increased to .34 for both genders. Substantial increments in reliability for this aggregation of predictor variables (rather than individual dimensions) were also seen.
Summary and Conclusions

In the Introduction, the legal and economic influences on work sample testing were acknowledged. It was observed that work sample assessment measures which are psychometrically sound, legally defensible, and economically viable are required. As a work sample test, the in-basket exercise is believed to enjoy the advantages of reduced bias and reduced adverse impact, although these attributes have not (yet) been empirically demonstrated. However, the in-basket also suffers the disadvantage of reduced utility because of its high development and administration costs, and the considerable time required both to train scorers and for trained scorers to evaluate a completed in-basket. Thus, it has been viewed as a costly component of managerial assessment. A unique challenge, then, facing in-basket developers has been in meeting the concerns of company executives to improve the economic efficiency of administrative ability assessment in an increasingly competitive corporate market.

Accordingly, the focus of the present research has been on the scoring of the in-basket exercise, long recognized as the weak link in its practical implementation. In fact, nearly forty years after the initial development of the in-basket exercise, Brannick et al. (1989) cited comparisons of scoring systems in in-basket research as a principal area for much-needed investigation. Excessive scoring time demands stemming from the complex and often subjective nature of in-basket scoring have limited the wide-scale application of high-fidelity, wideband exercises with large numbers of candidates for selection purposes.

It should be recognized that, because the purpose of measurement in this study is to make predictions about administrative ability, the most appropriate criterion-validation approach is to conduct a predictive validation study using a longitudinal research design. However, due to budgetary and time constraints, a predictive validation study was not
feasible, and so a concurrent validation design was substituted. Although such substitution is commonly made (Thornton & Byham, 1982), it is no doubt preferable to more closely match the purpose of measurement with study design.

It is also important to recognize the inherent limitations in accessing and measuring managerial performance given its multi-faceted nature and the lack of clear, consistent definitions and descriptions of managerial behaviour. Moreover, as noted in the Literature Review, problems of judgement bias remain a serious obstacle in performance appraisal assessment (Cascio, 1987). This study attempts to minimize this possible source of inaccuracy in the assessment of in-basket validity with the use of a factor analytically-derived aggregate of BOS scales as the criterion measure, based on immediate supervisors' ratings of observed frequencies of specific managerial behaviours. Despite this effort to minimize inaccuracy due to subjectivity in criterion measurement, it is no doubt clear that some inaccuracy remains; it is thus impossible to know, with complete confidence, whether any one in-basket scoring key is correct.

Given these limitations, the findings of the present study contain several important implications for both the in-basket developer and the in-basket user. First, contrary to expectations, it appears that Quality of Judgement scoring key development based on more logical, rather than empirical, principles is warranted. Although an empirical approach involving criterion-related validity results in the construction of scoring keys may appear promising, as suggested by Meyer (1970) and the findings from Preliminary Study 2, the present research clearly shows that the cross-sample generalizability of an empirically-derived key is inadequate, at least at sample-size levels like those in the present study.

Instead, paralleling the findings of Stern, Stein, and Bloom (1956), it appears that a panel-derived key, based on pooled judgements of experienced industry personnel,
remains psychometrically adequate in cross-validation. Specifically, more finely-grained analyses of the panel-based approach indicated that (a) panel size can be relatively small (here, 11 members were sufficient), (b) panel judgements can be combined relatively simply (a "liberal", 3-point scale was effective), (c) a less comprehensive list of action elements is as useful as an expanded list, and, lastly, (d) panel-derived Quality of Judgement keys allow company-to-company generalizability (i.e., the origin of the panel and the company or dataset to which it is applied need not correspond). These principles of panel and scoring key construction, when applied by in-basket test developers, should substantially reduce the historically high development costs associated with in-basket exercises.

Secondly, findings from the present study also result in greater understanding of the importance which should be afforded the stylistic dimensions of in-basket assessment. In summary, it was learned that, unlike the performance dimensions, the stylistic dimensions, when used in isolation, contribute little useful information to the prediction of administrative ability. However, previous research with the present exercise indicates that, when used in a linear composite along with performance dimensions (also involving a new Understanding of Situation measure), the stylistic dimensions contribute to a more global measure possessing improved reliability and criterion-related validity (Hakstian & Harlos, 1992). It should also be made clear that, on their own, stylistic dimensions do contribute useful diagnostic information about administrative performance.

A final consideration for the in-basket developer follows from the minimal shrinkage which occurred on cross-validation of the reduced item-subsets. It appears that, in conjunction with the principles of panel construction discussed above, action element-based scoring keys measuring performance on 8, rather than 21 items, result in adequate reliability and criterion-related validity for both the initial development and cross-validation samples. In conclusion, scoring-item reduction represents a new and
highly effective way to significantly improve the training and scoring efficiency of the high-fidelity in-basket exercise, thereby allowing more wide-scale applications of the exercise with large numbers of candidates for both training and selection purposes.
REFERENCES

Albemarle Paper Company v. Moody, 45 L. Ed. 2d 180 (U. S. Supreme Court 1975).


Appendix A

Listing of Selected Cognitive Ability and Personality Measures used in the Construct Validity Analysis

Cognitive Ability Measures:

Wonderlic Personnel Test
Culture Fair Intelligence Test (Scale 3)
Nelson-Denny Reading Test (Form E)
  Reading Speed
  Reading Comprehension
  Vocabulary
Flanagan Industrial Test
  Expression (Form A)
  Arithmetic (Form A)
Comprehensive Ability Battery (CAB)
  Ideational Fluency (Fi)
  Spontaneous Flexibility (Fs)

Aggregates:

General Intellectual Level: average of scores of Wonderlic Personnel Test and Culture Fair Intelligence Test
Processing Written Information: aggregate of Nelson-Denny Reading Test and Flanagan Expression Test
Cognitive Flexibility: aggregate of Ideational Fluency and Spontaneous Flexibility
Personality/Motivational Inventories:

California Psychological Inventory (revised)

Sixteen Personality Factor Questionnaire