

## INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.
2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.
3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.
4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.
5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

**University  
Microfilms  
International**

300 N. Zeeb Road  
Ann Arbor, MI 48106



8305971

Ariyo, Ademola

A SIGNAL DETECTION ANALYSIS OF AUDITORS' ANALYTICAL REVIEW  
JUDGMENTS

*The University of Arizona*

PH.D. 1982

University  
Microfilms  
International 300 N. Zeeb Road, Ann Arbor, MI 48106



A SIGNAL DETECTION ANALYSIS OF AUDITORS'  
ANALYTICAL REVIEW JUDGMENTS

by  
Ademola Ariyo

---

A Dissertation Submitted to the Faculty of the  
DEPARTMENT OF ACCOUNTING  
In Partial Fulfillment of the Requirements  
For the Degree of  
DOCTOR OF PHILOSOPHY  
WITH A MAJOR IN BUSINESS ADMINISTRATION  
In the Graduate College  
THE UNIVERSITY OF ARIZONA

1 9 8 2

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read  
the dissertation prepared by ADEMOLA ARIYO

entitled A SIGNAL DETECTION ANALYSIS OF AUDITORS'

ANALYTICAL REVIEW JUDGMENTS

and recommend that it be accepted as fulfilling the dissertation requirement  
for the Degree of DOCTOR OF PHILOSOPHY.

Tia Soloman

December 17, 1982

Date

Wm. S. Watts

December 20, 1982

Date

Walter R. French

20 December '82

Date

Don Vickers

20 November 1982

Date

X Ernest D. Summers

Y 20 December 1982

Date

Final approval and acceptance of this dissertation is contingent upon the  
candidate's submission of the final copy of the dissertation to the Graduate  
College.

I hereby certify that I have read this dissertation prepared under my  
direction and recommend that it be accepted as fulfilling the dissertation  
requirement.

Tia Soloman  
Dissertation Director

December 20, 1982

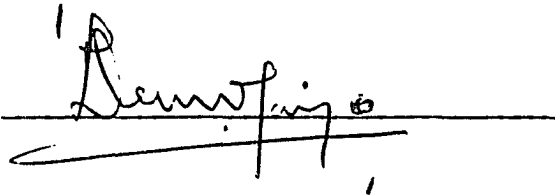
Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: \_\_\_\_\_

A handwritten signature in dark ink, appearing to read "L. Murphy", is written over a horizontal line. Below the line, there is a long, sweeping horizontal stroke that extends to the left and then curves back towards the right.

## ACKNOWLEDGMENTS

I would like to thank the following members of my dissertation committee whose interests in my educational progress go beyond the desire to roll out another "product" through the doctoral degree mill: Dr. Ira Solomon (Chairman), Dr. William R. Ferrell, Dr. George W. Summers, Dr. Don Vickrey, and Dr. William Waller. Although all the committee members made significant contributions towards this dissertation, I like to give special recognition to the continuous advice and encouragement I received from both Dr. Solomon and Dr. Ferrell throughout my stay at the University of Arizona. I also would like to thank the following: The Rockefeller Foundation for its financial support throughout the period of my graduate studies; the head and faculty of the Department of Economics, University of Ibadan in Nigeria, for their moral support and unwavering interest in my academic progress; The University of Arizona Foundation for a Graduate Academic Scholarship Award; the public accounting firms which provided both the data and the subjects for my research study; the Department of Accounting and the Department of Systems and Industrial



Engineering of the University of Arizona, for research-related financial and technical assistance.

In addition, I thank Dr. Russell M. Barefield, the head of the Department of Accounting, University of Arizona, for providing an environment conducive to the early completion of my doctoral studies; Dr. John W. Dickson for his interest and guidance both as my minor program advisor and a teacher in research methodology, and Ernest Yaw Baafi, a colleague who, without charge, spent numerous days and nights writing and debugging the computer programs required for my data analysis.

My love and gratitude go to my children, Adetola and Adeola, for whom adequate fatherly attention at the most formative stages of their lives was a luxury, and to my wife, Olamide, whose love and understanding endured.

Finally, I dedicate this dissertation to my mother, Obi, the architect of whatever I am.

## TABLE OF CONTENTS

	Page
LIST OF ILLUSTRATIONS . . . . .	ix
LIST OF TABLES . . . . .	xi
ABSTRACT . . . . .	xii
CHAPTER	
1. INTRODUCTION . . . . .	1
Background to this Study . . . . .	4
Purpose of the Study . . . . .	8
Overview of the Dissertation . . . . .	11
2. THE AUDIT DECISION PROCESS . . . . .	14
The Audit Task . . . . .	14
The Nature of Analytical Review . . . . .	18
Analytical Review and Audit Planning . . . . .	20
Analytical Review and Non-Audit Services . . . . .	26
3. THE ANALYTICAL REVIEW LITERATURE. . . . .	28
Statistical Analysis Review (SAR). . . . .	29
Judgmental Analysis Review (JAR) . . . . .	34
The Need for a Study of Auditor Judgment in AR. . . . .	36
Research Issues: An Outline . . . . .	40
Research Issue Number One:	
Detectability of Auditor Judgments in PAR. . . . .	41
Research Issue Number Two:	
Implicit Loss Functions Affecting Auditors' PAR Judgment . . . . .	41
Research Issue Number Three:	
Effect of Nature of Task on Auditors' PAR Judgment . . . . .	43
Research Issue Number Four:	
Effect of State of AIC on Auditors' PAR Judgments. . . . .	44
Research Issue Number Five:	
Effect of Functional Level on Auditors' PAR Judgment. . . . .	45

TABLE OF CONTENTS--Continued

	Page
Research Issue Number Six:	
Auditors' Sensitivity to Their	
Degree of Uncertainty. . . . .	47
Research Issue Number Seven:	
Information Items Required by Auditors	
to Facilitate Their PAR Judgments. . . . .	48
4. SIGNAL DETECTION ANALYSIS. . . . .	59
Signal Detection Tasks . . . . .	60
Signal Detection Theory. . . . .	61
Relative Operating Characteristics . . . . .	65
Indices of Detectability . . . . .	68
Experimental Approaches in SDT . . . . .	71
"Yes-No" Mode. . . . .	71
Two-Alternative Forced Choice Mode . . . . .	73
Rating Scale Mode . . . . .	74
The Decision Variable Partition Model. . . . .	77
General Description of the Model . . . . .	77
Independence of Knowing and Knowing	
That One Knows . . . . .	80
Calibration and DVPM . . . . .	81
The Criterion. . . . .	83
Calculating the Criterion. . . . .	88
PAR as a Signal Detection Task . . . . .	89
5. METHODOLOGY. . . . .	94
Scope of the Study . . . . .	94
The Subjects . . . . .	97
The Case Studies . . . . .	99
Criteria for Selecting and Classifying Account Items . . . . .	101
Experimental Task. . . . .	105
Conclusions Versus Decisions in Experimental	
Tasks: A Distinction. . . . .	107
The Pilot Study. . . . .	110
Administration of the Study. . . . .	110
Research Issues Addressed. . . . .	112
Research Issue Number One: What is the	
Detectability of Auditors' Judgments?. . . . .	112
Research Issue Number Two: What Type of	
Responses Biases Are Exhibited by the Auditor? . . . . .	114
Research Issue Number Three: What is the	
Effect of Task Criterion on Auditors'	
Performance. . . . .	117

TABLE OF CONTENTS--Continued

	Page
Research Issue Number Four: What is the effect of the state of AIC on Auditors' performance? . . .	118
Research Issue Number Five: What is the effect of functional level on auditors' performance? . . .	119
Research Issue Number Six: How Effective Are Auditors at Communicating Their Knowledge in Preliminary Analytical Review Tasks? . . . . .	120
Research Issue Number Seven: What Types of information do auditors require for PAR judgment? . . . . .	123
Applicability of SDT to Groups of Subjects . . . . .	126
 6. EFFECTS OF PRIOR SIGNAL PROBABILITY, NUMBER OF TRIALS, AND POOLING OF RESPONSES ON DETECTABILITY. . . . .	 130
Prior Signal Probability . . . . .	132
Number of Trials . . . . .	134
Pooling of Responses . . . . .	137
Simulation Experiment. . . . .	138
Analysis and Discussion of Results . . . . .	140
Comment on the Simulation Results. . . . .	145
 7. THE DATA ANALYSIS. . . . .	 148
Discussion of Results by Research Issue . . . . .	148
Research Issue Number One: Detectability of Auditors' Responses. . . . .	148
Research Issue Number Two: Subjects' Judgment Bias . . . . .	151
Research Issue Number Three: Effect of Task criterion on Auditors' Performance. . . . .	156
Research Issue Number Four: The Effect of Internal Control on Auditors' Responses. . . . .	167
Research Issue Number Five: The Effect of Functional Level on Auditors' Performance. . . . .	171
Research Issue Number Six: Calibration of Subjects' Responses. . . . .	174
Research Issue Number Seven: Information Required for PAR Judgments . . . . .	187

TABLE OF CONTENTS--Continued

	Page
8. SUMMARY, LIMITATIONS, IMPLICATIONS OF RESEARCH FINDINGS, AND SUGGESTIONS FOR FURTHER RESEARCH. . . . .	194
The Limitations of the Study. . . . .	194
The Major Findings of the Study . . . . .	197
Research Issue One. . . . .	197
Research Issue Two. . . . .	198
Research Issue Three. . . . .	198
Research Issue Four . . . . .	199
Research Issue Five . . . . .	199
Research Issue Six. . . . .	199
Research Issue Seven. . . . .	199
Implications of Research Findings and Suggestions for Further Research. . . . .	200
APPENDIX A: EXPERIMENTAL MATERIALS . . . . .	206
APPENDIX B: AN OVERVIEW OF GREY & MORGAN'S MODEL AND THE PROCEDURES FOR ANALYZING THE STUDY'S DATA. . . . .	228
BIBLIOGRAPHY. . . . .	233

## LIST OF ILLUSTRATIONS

Figure	Page
1-1 Experimental Design. . . . .	10
2-1 The Role of Analytical Review in Audit Planning. . . . .	25
3-1 Example of Calibration Curves for a Two-Alternative, Forced-Choice Task . . . . .	51
4-1 Example of a Decision-Making Matrix. . . . .	62
4-2 Underlying Distribution for SN, N. . . . .	64
4-3 The ROC Graph. . . . .	66
4-4 ROC Curve as a Measure Detectability . . . . .	72
4-5 The Partition on $X = Y$ Which Gives Perfect Calibration for a YN(FR) Task When the Conditional Distributions of Y are Normal, Unit Variance with $d' = 2.5$ and $P(c) = .5$ . . .	79
4-6 Signal Detection Theory (SDT) and Preliminary Analytical Review (PAR): A Comparison of Concepts. . . . .	93
5-1 Experimental Design. . . . .	95
5-2 Categorization of Account Item as N or SN by Account Classification Approach. . . . .	106
6-1 Effect of $P(SN)$ and Number of Observations (NT) on the Underestimation of Area Under the ROC Curve. . . . .	144
7-1 Overall Calibration of Subjects' Responses by Account Classification Criterion . . . . .	178
7-2 Account Classification Criterion I: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses . . . . .	179
7-3 Account Classification Criterion II: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses . . . . .	179

LIST OF ILLUSTRATIONS--Continued

Figure	Page
7-4 Account Classification Criterion III: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses . . . . .	.180
7-5 Account Classification Criterion I: Effect of Funtional Level on Calibration of Subjects' Response.. . . .	.181
7-6 Account Classification Criterion II: Effect of Functional Level on Calibration of Subjects' Responses. . . . .	.182
7-7 Account Classification Criterion III: Effect of Functional Level on Calibration of Subjects' Responses. . . . .	.183
AB-1 The Symmetric Probability Scale Used to Code the Subject's Responses. . . . .	.231
AB-2 Response Categories for Coding the Subjects' Responses.. .	.231

## LIST OF TABLES

Table	Page
6-1 Simulated Values of Area Under the ROC Curve. . . . .	141
6-2 Two-Way Analysis of Variance: Effects of Number of Trials (NT) and Prior Signal Probability P(SN) on Area Under the Curve. . . . .	142
6-3 Effects of Number of Trials (NT) and Prior Signal Probability P(SN) on Variability of Area Under the Curve. .	146
7-1 Area Under the ROC Curve. . . . .	149
7-2 Effect of Pooling of Responses on Area Under the ROC Curve.	152
7-3 Index of Response Bias by Account Classification Approach..	154
7-4 Characteristics of Auditor's Responses. . . . .	158
7-5 An Example of the Procedure for Evaluating the Effect of Task Criterion on the Subject's Responses . . . . .	160
7-6 Test for Differences in Characteristics of Auditors' Responses by Task Criterion . . . . .	163
7-7 Effect of Task Criterion on Judgment Bias . . . . .	166
7-8 Effect of Quality of Internal Control on Auditor's Responses . . . . .	170
7-9 Effect of Functional Level on Auditors' Performance . . . .	173
7-10 Calibration of Subjects' Responses. . . . .	175
7-11 Information Items Required for PAR Judgments. . . . .	188
7-12 Auditor's Overall Degree of Consensus Regarding the Relative Importance of Information Items Required for PAR Judgments . . . . .	192



## ABSTRACT

The auditors' preliminary analytical review procedures (PARPs) have recently received increased attention in the accounting literature, because of the growing realization that PARPs may significantly enhance audit effectiveness and efficiency. Although both judgmental and statistical PARPs (JARPs and SARPs respectively) are recognized in the professional literature, earlier research has concentrated exclusively on statistical ARPs (SARPs). This research bias is inappropriate, given that SARPs merely supplement, but do not replace, auditor-judgments in PARPs. Enhancement of audit effectiveness and efficiency requires, therefore, evidence bearing on several aspects of auditors' PARPs judgments. This dissertation uses a model based on Signal Detection Theory to provide evidence relating to the following aspects of auditors' PARPs judgments: (a) judgmental accuracy, (b) decision errors, (c) implicit loss functions, and (d) information required to facilitate PARPs judgments.

The major findings of the study were: (1) auditors can make reasonably accurate AR judgments on the basis of limited information available at the onset of an

audit; (2) their responses were affected by judgmental biases, with a propensity to flag for intensive audit account book values which are fairly presented; (3) the auditors' judgments were miscalibrated, being mostly overconfident; and (4) simple ARPs such as ratio analysis, scanning, and comparisons amongst data, are those preferred by auditors for their PARPs judgments.

This study's findings suggest the need to identify the causes, and subsequently mitigate the effects, of judgment biases before the potential of auditors' AR judgments at enhancing audit effectiveness and efficiency can be fully realized.

## CHAPTER I

### INTRODUCTION

The ultimate objective of the external auditor is to express an opinion on the conformance of management's financial statement representations with generally accepted accounting principles (GAAP). The auditor chooses account items for investigation, on a select basis, as support for his/her opinion. Such audit evidence provides reasonable protection against two types of risks: (a) the risk that material errors will occur in the accounting process by which the financial statements are developed, and (b) the risk that any material errors that should occur will not be detected by the auditor's examinations. It is incumbent on the auditor, therefore, to employ an audit approach which maximizes the chance of selecting for audit those account items that are most likely to be materially misstated.

To enable him/her to obtain the evidence required for this protection, generally accepted auditing standards (GAAS) suggest that the auditor (a) perform a proper study

and evaluation of the existing internal control system, and (b) obtain sufficient, competent evidential matter, through substantive tests of the accounts. These substantive tests are (a) tests of details of transactions and balances, and (b) analytical review procedures. Although the choice of a specific audit procedure is a matter of professional judgment, GAAS require that effectiveness be the overriding criterion guiding this choice.

The goal of enhancing audit effectiveness and efficiency has stimulated renewed interest in analytical review (AR) procedures, to the extent that some observers (e.g., Biggs, 1981) have speculated that the audit of the future might consist only of two elements: control reviews and analytical reviews. Two broad approaches to analytical review (AR) can be distinguished: judgmental analytical review (JAR) and statistical analytical review (SAR). JAR is characterized by the use of insight, experience, knowledge of the client firm's specific environmental data, and professional judgment by the auditor to determine (a) a reasonable range of values for the account balance, and (b) to evaluate the significance of any difference from the account book value. SAR, on the other hand, uses formal economic and statistical models to relate the account balance to environmental

variables and related account items, as a basis for determining the reasonableness of reported book values.

Two roles of AR also have been identified in the professional literature: (a) preliminary AR (PAR) and (b) substantive AR. The former usually is applied at the onset of an audit to assist in identification of account book values which are likely to be materially misstated and, consequently, ones to which more of the available audit resources should be devoted. This AR role has been described as the "attention-directing" role of AR (Kinney and Felix, 1980). The latter typically is employed either (a) during the conduct of the audit in conjunction with other audit procedures (i.e., as a test-of-details substitute [Kinney and Felix, 1980]), or (b) at or near the end of an audit as an overall review of the financial information.

In pursuit of the goal of enhancing audit efficiency and effectiveness, prior research studies (e.g., Akresh and Wallace, 1980; Kinney, 1978; Neter, 1980; and Stringer, 1975) have focused on the attention-directing role of AR. Furthermore, previous studies have concentrated only on SAR, perhaps stemming from the belief that SAR is more objective than JAR (see, for example, Akresh and Wallace, 1980; Stringer, 1975, for further details). Consequently, the role of auditor judgment in AR has largely been ignored.

A notable exception is the study by Blocher, Esposito, and Willingham (1981), in which the authors evaluated certain situational and individual variables on auditor judgment in AR procedures for a payroll audit. Their findings suggest that the perceived role of AR at the planning stage of an audit is likely to be subject to more inconsistent auditor judgments than its perceived role at the usage (substantive testing) phase. Blocher, et al's (1981) study, however, did not address directly the effect of several aspects of auditor's AR judgments on audit effectiveness and efficiency.

#### Background to This Study

I consider the concentration of research efforts exclusively on SAR inappropriate in view of the following. First, in practice, the formulation, implementation, and interpretation of SAR require auditor judgments. In fact, the results of SAR merely supplement, and do not substitute for, auditors' AR judgments. The effectiveness of SAR ultimately depends, therefore, on the nature and characteristics of auditor judgments. Perhaps it is in recognition of this fact that some observers (e.g., Mock, Biggs, and Watkins, 1982) have suggested that research directed at understanding the auditors' judgment process in AR is required before appropriate statistical models can be developed. Second, many public accounting (i.e.,

CPA) firms employ only JAR. Due to cost considerations, SAR typically is applied only to "large" client engagements.

Third, some researchers (e.g., Kinney and Felix, 1980, p. 102) have indicated that the so-called objective (i.e., statistical) models necessarily use only a small part of the information which may be available to the auditor. Hence, they have suggested that research be conducted into the use of expert judgments in AR. It appears, therefore, that any realistic effort aimed at enhancing audit effectiveness requires, at the minimum, evidence regarding the characteristics of auditors' AR judgments. Chapter 3 of this study, therefore, provides a discussion of the need to study auditor judgment in preliminary analytical review (PAR).

The validity of the results of any study designed to provide evidence on chosen characteristics of auditor judgments in PAR depends, to a large extent, on the appropriateness of the statistical and/or research techniques employed. Most of the earlier studies of auditor judgments, especially in the internal control context (e.g., Ashton, 1974) have employed measures such as correlation coefficients to evaluate auditor judgments. Whether these measures are appropriate for the experimental tasks is, however, still an unresolved issue, given the criticisms and limitations noted in the

literature about their usage for evaluating subjective judgments (for example, see Birnbaum, 1973; Remus and Jenicke, 1978). Furthermore, these studies simply concluded that there was a great variability in auditor judgments with little or no attempt to investigate the underlying causes.

In this dissertation, I use a statistical model which I consider to be most appropriate for analyzing auditors' PAR judgments. This belief is based upon the following logic. To make a PAR judgment, the auditor evaluates the cues available to him/her at the onset of the audit. Using his/her prior knowledge regarding the possible co-occurrence of the cues and the characteristics of two audit populations (i.e., fairly presented book values and materially misstated book values), the judge (i.e., auditor) decides from which population the book value under consideration comes. If the auditor decides that the book value is materially misstated, that indicates that s/he believes the cues contain a signal (imbedded in noise). Otherwise, s/he believes that the uncertain relationship between the cues and the possible deviation of the book value from expectation is merely due to chance (noise) alone. PAR, therefore, is essentially a detection task, for which the principles of signal detection theory seem applicable.



Ferrell and McGoey (1980) have developed a model for representing judgments in such detection tasks, which is suitable for the present study. The model, called the Decision Variable Partition Model, breaks the judgmental tasks into two aspects: the detection aspect, and the probability encoding aspect. The detection aspect provides evidence regarding the accuracy of subjects' judgments independent of the motivational factors (e.g., implicit loss functions) which may affect the subject's judgments; the probability encoding aspect enables one to determine the calibration of the subjects' judgments. The details of the model are presented in Chapter 4.

The accounting literature indicates that auditors' PAR judgments can be affected by factors such as the perceived quality of the internal control and the auditor's functional level. This study provides evidence bearing on the likely effects of these factors on auditor judgments in PAR.

The accounting literature also indicates that differences in information search behavior can affect judgment accuracy and variability in experimental tasks. In addition, some researchers have suggested that the types of information which auditors require for their PAR judgments should be identified before appropriate statistical models can be developed. Hence, in this study, I request the auditor-subjects to indicate, in a

decreasing order of importance, the information items they would have required to facilitate their AR judgments in practice, regardless of those contained in the Sssq experimental materials.

The background for this study can be summarized as follows. First, no study has been reported bearing on the characteristics of auditor judgment in PAR. Second, most of the previous studies have used statistical techniques with little or no regard to the appropriateness of such techniques for evaluating judgments under uncertainty. This study employs a signal detection model which is appropriate for evaluating auditor judgments in PAR tasks. Third, earlier studies merely revealed the observed variability in auditor judgments without identifying some of the potential causes of the observed results. In this study, I provide evidence regarding the possible causes (e.g., implicit loss functions) of the features of auditors' PAR judgments.

#### Purpose of the Study

My aim is to provide evidence bearing on several aspects of auditor judgments in PAR tasks, and to suggest the implications of the evidence from the perspective of audit effectiveness and efficiency. Specifically, I investigate (a) auditors' detectability, which provides a

measure of the extent to which auditors can identify, on the basis of limited information available at the onset of an audit, account items which are materially misstated; (b) the implicit loss functions employed and the type of decision errors auditors have a propensity to commit; (c) auditors' degree of sensitivity to their level of uncertainty in PAR tasks; (d) the effect of functional level and other environmental factors, such as the state of internal control, on auditors' AR judgments, and (e) the information items which auditors consider relevant to the formulation of PAR judgments. I performed an experiment in which practicing auditors provided responses to a set of questions. The responses were used to generate the data bearing on the above issues which were investigated.

Each participant in the study was provided with two experimental cases: (1) a case in which the internal control is adjudged strong, and (2) a case in which the internal control is adjudged relatively weak. The participants' responses were evaluated overall, by state of internal control, and by functional level. This process led to the research design indicated in Figure 1-1.

The major findings of the research study are:

1) The detectability aspect of auditors' PAR

judgments is reasonably high, given the constraining

		INTERNAL CONTROL	
		WEAK	STRONG
FUNCTIONAL LEVEL	SENIOR		
	MANAGER		

Fig. 1-1. Experimental Design

- factors inherent in the experimental task. This result indicates that auditors can make fairly accurate judgments on the basis of limited information.
- 2) The auditors' responses were affected by judgmental biases. The predominant form of bias is the tendency to flag for intensive audit account book values which are fairly presented. This strategy suggests that auditors may be risk-averse when making audit decisions.
  - 3) The auditor-subjects' responses were miscalibrated, the responses being mostly overconfident. There was, however, no significant effect of AIC and functional level on the miscalibration of the subjects' responses.
  - 4) Consistent with findings reported in earlier studies, evidence indicates that simple AR procedures such as ratio analysis, scanning, and comparisons amongst data, were the ones most often indicated by the auditor-subjects as being useful for formulating PAR judgments.

#### Overview of the Dissertation

The remainder of the dissertation is divided into seven chapters. Chapter 2 discusses the audit decision process as it relates to the objectives of this study. In

particular, I discuss (a) the audit task, (b) the nature of AR, and (c) the role of AR in both the audit process and other services which auditors provide to their clients.

In Chapter 3, I provide a review of the AR literature and note that AR research efforts largely have been concentrated exclusively on SAR to the neglect of JAR. I then justify the need for a study of auditor judgment in the PAR context by indicating that SAR does not preempt JAR, and that JAR will always continue to be used in practice in the foreseeable future. Thereafter, I introduce the research issues addressed in this study.

To provide a justification for the use of a statistical method based on the signal detection theory for analyzing this study's data, I describe the concept of calibration which has been used in prior studies to address some of the research issues (e.g., the concept of knowing that one knows) investigated herein. The shortcomings of using calibration only for evaluating judgments under uncertainty are then highlighted. Finally, I discuss the need to provide evidence regarding the calibration of auditors' PAR judgments, because of its relevance to audit effectiveness and efficiency concerns.

Chapter 4 discusses signal detection theory and its relevance to PAR, and also the Decision Variable Partition Model developed by Ferrell and McGoey (1980)

which is used in this study. This is followed in Chapter 5 by a description of the methodology of this study, including a discussion of the specific research issues investigated.

Chapter 6 reports the results of a simulation study designed to provide evidence regarding the effect on observer detectability of (a) prior signal probability, (b) number of stimulus observations (trials), and (c) pooling of responses. The results provide a basis for developing expectations regarding the likely effects of these factors on the performance of the auditor-subjects of this study. This is followed by a description of the data analysis in Chapter 7, while Chapter 8 presents the conclusions, implications, and suggestions for further research.

## CHAPTER 2

### THE AUDIT DECISION PROCESS

#### The Audit Task

The external auditor's ultimate objective is to determine whether the financial statements of a business firm "present fairly" its financial position, results of operations, and changes in financial position, in conformity with generally accepted accounting principles (Statement of Auditing Standards [SAS] No. 1, 1979). Operationally, the auditor and those who rely on his/her opinion require a reasonable protection against two separate risks: (a) the risk that material errors will occur in the accounting process by which the financial statements are developed, and (b) the risk that any material errors that should occur will not be detected in the auditor's examination (SAS No. 1, Sec. 320a.14).

The professional literature indicates that the probability that the first type of risk will occur is inversely related to the strength of the accounting internal control (AIC) system. The auditor, therefore,



has no control over the chances of the occurrence of this type of risk, since it is dependent on the data processing system for which management is responsible. Auditing theory, therefore, conservatively assumes that there is a 100% probability that the first type of risk will occur.

Hence, to enable the auditor to make a judgment regarding the reliability of the AIC, generally accepted accounting standards (GAAS) suggest that there be

a proper study and evaluation of the existing internal control as a basis for reliance thereon and for the determination of the resultant extent of the tests to which auditing procedures are to be restricted (SAS No. 1, Sec. 320.01).

This recommendation is based on the assumption that the existence of a satisfactory accounting internal control (AIC) reduces the probability that material errors in the accounts will occur and go undetected. However, since the AIC system is not perfect, the auditor cannot place complete reliance on AIC as the only means of gathering the relevant audit evidence. The extent of reliance to place on the AIC is, therefore, determined by the auditor after an evaluation and study of the AIC. If we denote by  $C$  the degree of reliance the auditor places upon a given AIC, then  $(1-C)$  denotes the risk that material errors will go undetected through the AIC.

Given the inherent imperfection of the AIC, the third standard of field work suggest that, to reduce the second type of risk,

sufficient competent evidential matter is to be obtained through inspection, observation, inquiries, and confirmations to afford a reasonable basis for an opinion regarding the financial statements under examination (SAS No. 1, Sec. 320.69).

SAS No. 1, (Sec. 320.70) notes that the evidential matter required by the third standard of field work can be obtained through (a) tests of details of transactions and balances [TDTB], and (b) analytical review [ARP]. Both (a) and (b) are referred to as substantive tests in the professional literature. The auditor also decides upon the degree of reliance to place upon these substantive tests, say,  $\underline{S}$ , on the basis of information derived from applying these procedures. Since the substantive tests comprise TDTB and ARP, one can decompose  $\underline{S}$  into the degree of reliance placed upon (a) TDTB, say,  $\underline{T}$ , and (b) ARP, say  $\underline{A}$ . Therefore,  $(1-T)(1-A)$  represent the risk that material errors which occur will not be detected by the substantive tests.

The multiplicative effect of the risk that material errors which occur will not be detected by the combined application of these procedures is referred to as the ultimate risk (UR), the minimization of which is the

GAAS indicate that, in choosing an audit procedure or a combination of audit procedures, effectiveness necessarily is the overriding consideration (SAS No. 1, Sec. 320.73, 1979, emphasis added). GAAS also suggest that efficiency is an appropriate consideration in choosing between procedures of similar effectiveness. It is noteworthy, however, that the study and evaluation of AIC is the only audit procedure available for reducing the first type of risk. Hence, a consideration of the choice of the most effective audit procedures actually relates to a choice of the most effective combination of the substantive tests to reduce the second type of risk.

Before discussing the need for a study of auditor judgments in AR, it appears reasonable to understand the nature and the importance of AR, and its potential usefulness as a means of enhancing audit effectiveness and efficiency. Hence, in the following sections I discuss (a) the nature of analytical review, and (b) the audit framework and the role of analytical review in the planning stages of an audit and other nonaudit services which accountants provide to client firms.

#### The Nature of Analytical Review

SAS No. 23 (1981) states that "analytical review procedures are substantive tests of financial information made by a comparison of relationships among data." It also

auditor's main objective. Specifically, UR is defined as follows:

$$UR = K(1-C)(1-T)(1-A) \quad (2-1)$$

where K refers to the risk that material errors will occur in the accounting process by which the financial statements are developed. However, since GAAS assumes K to be equal to 1, then

$$UR = (1-C)(1-T)(1-A) \quad (2-2)$$

Equation (2-2) indicates that UR is inversely related to the degree of reliability placed upon each of the audit procedures discussed above.

The second standard of field work also recognizes that the extent of substantive tests required to constitute sufficient evidential matter under the third standard may properly vary inversely with the auditor's reliance on AIC (SAS No. 1, Sec. 320A.19). It also recognizes that, regardless of the extent of reliance on AIC, the auditor's reliance on the substantive tests may be derived from tests of details, from ARP, or from any combination of both, as the auditor deems appropriate in the circumstances. Hence, to enable the auditor to make use of the framework in equation (2-2), knowledge of the usefulness of AR, especially in the attention-directing mode, is required.

states that a basic premise underlying the application of ARPs is that relationships among data may reasonably be expected by the auditor to exist and continue in the absence of known conditions to the contrary.

The guiding framework for the evaluation of the results of AR is the auditor's expectations of what the recorded amounts should be, or what the range of possible values might reasonably be expected to be. The acceptable degree of variation from expectation would depend, for example, on other evidence, on the reasonableness of the identified relationships, on costs and, possibly, on the auditor's attitude toward risk. The determination of what constitutes "unusual" or "out-of-line" account items is, therefore, a matter of professional judgment.

There are two uses of AR described by SAS No. 23: the AR may indicate (a) the need for additional procedures, or (b) that the extent of other auditing procedures may be reduced. Kinney and Felix (1980) have described these as "attention-directing" and "tests-of-details substitute" uses of AR respectively. These two uses of AR are, however, related to one another. For example, the higher the perceived effectiveness of an ARP as an attention-directing tool, the more reliance will be placed on such a procedure and, hence, the less other substantive tests will be required.

Corresponding to the varying objectives of AR, SAS No. 23 lists the following combinations of timing and objectives:

- a) In the initial planning stages to assist in determining the nature, extent and timing of auditing procedures by identifying, among other things, significant matters that require consideration during the examination.
- b) During the conduct of the examination in conjunction with other procedures applied by the auditor to individual elements of financial information.
- c) At or near the conclusion of the examination as an overall review of the financial information (para. 5).

AR performed at the initial planning stage of an audit is referred to as preliminary AR (PAR), while the AR performed at or near the conclusion of the audit is referred to as substantive AR.

#### Analytical Review and Audit Planning

AR's ability to enhance audit effectiveness and efficiency depends on its usefulness as an audit planning tool. Hence, in this section, I discuss the potential usefulness of AR in the audit planning process, as suggested by both the academic and professional literatures.

Having accepted an audit engagement, prudence demands that the auditor plan and schedule the audit in the most efficient and effective manner. For example, the

auditor might wish to perform as much preliminary work as possible at an early date, rather than leave the entire task to the end of the period under audit. This requires that the auditor have a reasonable understanding of general business and industry conditions, and the client firm's accounting policies and procedures. Based on this background knowledge and the findings of the preliminary work performed, the auditor begins the engagement planning by scheduling and programming the activities to be performed at each stage of the audit.

Robertson (1979, p. 148) claimed that PAR plays a large role in this initial planning phase of the audit. It is useful for identifying potential problem areas in the financial statements and conditions that may require an extension and modification of anticipated audit procedures. Taylor and Glezen (1979) note that an understanding of the client firm's business operations allows the auditor to use audit techniques, such as analysis of operating and financial ratios, as more efficient audit tools. They also note that such an understanding facilitates detection of financial statement items that appear unusual. Hence, the auditor can be efficient in the sense of concentrating on the areas in which material misstatements are likely to be contained.

In essence, PAR can be viewed as a means of identifying at an early point in the audit areas that may

present problems or require special attention later on. In fact, Taylor and Glezen (1979) labelled such tests of reasonableness "predictive auditing."

Kinney (1981) also classifies PAR as part of the orientation stage of the audit process. He states that

A part of the orientation stage would be a study of the available book values and comparison of relationships between and among the book values and other data such as similar data from other firms or other time periods or other data such as budgets. These comparisons may yield book values which seem to be "out of line" with what one would expect. The comparisons give the auditor some basis for an assessment of the probability that material error is present. The orientation stage also gives the auditor a basis for beginning the study of the internal accounting control (p. 4).

The discussion above suggests that there is a high degree of consensus among researchers regarding the role of AR in the audit process. This consensus appears consistent with the role assigned to AR in professional practice. The methods currently being used by some CPA firms are discussed below to buttress this point.

One audit approach, developed by Peat, Marwick, Mitchell and Co. (PMM) is called the Systems Evaluation Approach (SEA). It clearly identifies a role for AR at the planning, interim, and final phases of the audit. At the planning stage, the SEA indicates that the review of financial data constitutes the initial information on which the planning of the other phases of the audit is



based. Its main purpose is to assist the auditor in identifying unusual relationships of financial data that may have audit significance, and to assist in determining the scope and relative emphasis of the audit work (Taylor and Glezen, 1979). The final phase also indicates, again, the important role of AR, which takes place as soon as the year-end information becomes available. As in the planning phase, the purpose of this review is to identify unusual relationships that may have audit significance, thereby assisting the auditor in determining whether any modification to the audit program is warranted by changes in trends and conditions subsequent to the interim phase of the audit.

The role of AR in SEA is identical with the one described by Touche Ross and Co. (1981) in its TRAP (Touche Ross Analytical Process) audit approach. In the TRAP, the audit is broken down into three phases, the first and the last of which are relevant to this discussion. Phase I, which is concerned with Planning and Evaluation, aims at assisting the auditor in identifying areas in which audit attention should be concentrated to satisfy standard audit objectives. Analytical comparisons both of financial and operating information, as well as planned control and other audit tests, are performed at this stage. Phase III is Completion of the Audit, when

analytical procedures again are used to assess the overall reasonableness of transactions and account balances.

Discussions with representatives of other national CPA firms indicate similar AR roles and procedures in their respective firms. The importance of AR as a planning tool in the audit process, as described by SEA and TRAP, and through discussions with auditors from other CPA firms, is consistent with the procedures described by Robertson (1979:146-150), and Kinney (1981). Hence, it appears that, both in theory and in practice, there is consensus regarding the role of AR in the audit process.

Of direct relevance to this study is the role of PAR in audit planning and the resulting allocation of audit efforts. One important factor which can have impact upon the decision of how to allocate audit efforts is the state of AIC. As indicated earlier, GAAS suggest that there should be an inverse relationship between the quality of IAC and the extent of substantive tests. In essence, SAS No. 1, Sec. 320 indicates that after the auditor reviews the system of IAC and performs compliance tests of the IAC on which s/he intends to place significance reliance, it is then appropriate to limit substantive tests (e.g., AR) in response to the presence of strong and effective IACs. The relationship between the state of AIC and the allocation of audit efforts is described in Figure 2-1.

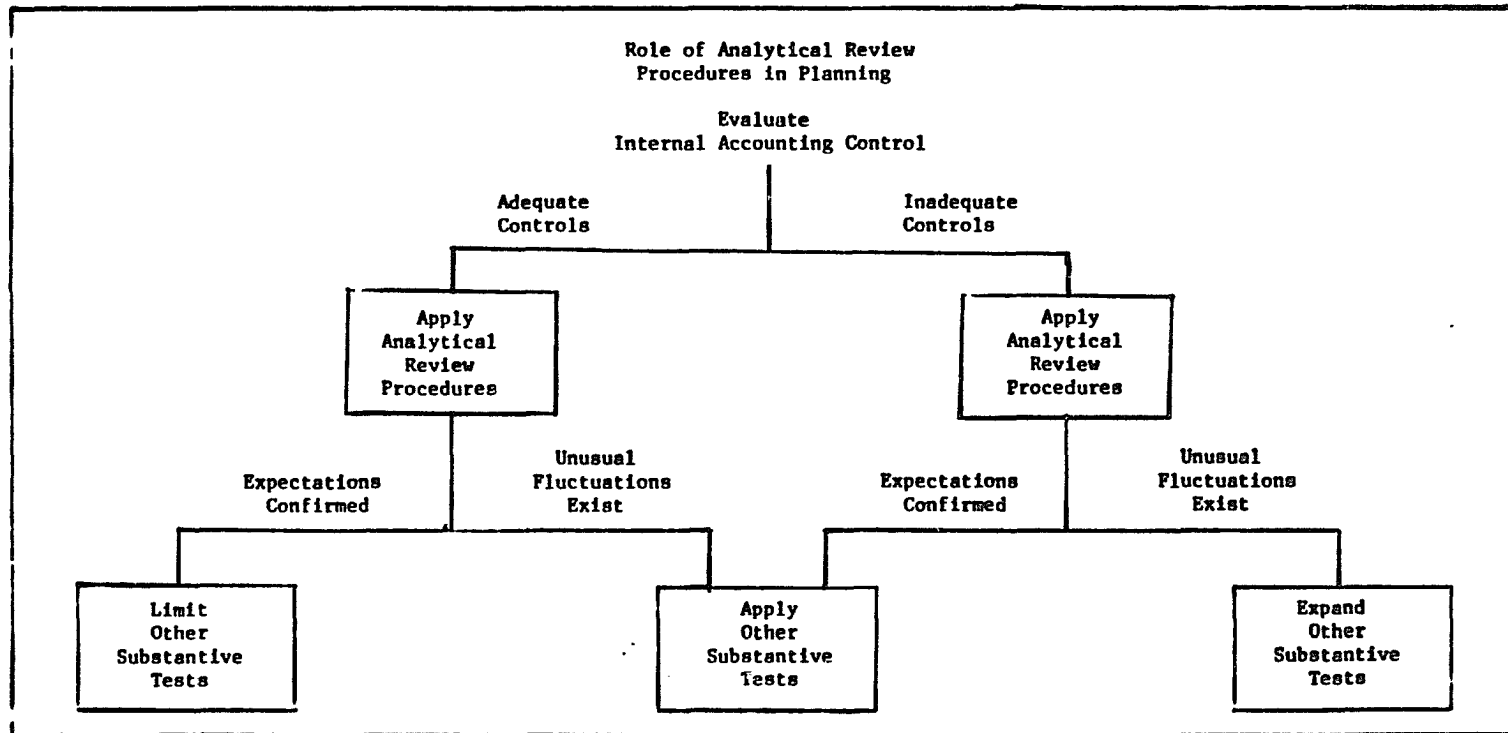


Fig. 2-1. The Role of Analytical Review in Audit Planning.

Source: Holder and Collmer (1980).

The figure indicates that when unusual fluctuations are found in a situation in which the AIC is adjudged adequate, the auditor will ordinarily apply other substantive tests. Similarly, when unusual fluctuations are found when the AIC is adjudged weak, the figure indicates that the auditor should expand other substantive tests. However, since an effective PAR should assist the auditor in identifying some of these unusual fluctuations at the onset of the audit, application of substantive tests can be reduced if either (a) no unusual fluctuations are found, or (b) PAR were performed although unusual fluctuations are found.

Analytical Review and  
Nonaudit Service

AR also can be (and is being) used to facilitate the performance of other services which accountants provide for client firms. Specifically, AR presently is used during (a) reviews of Interim Financial Information (SAS No. 36) and (b) Reviews of Financial Statements of non-public entities (Statements of Standards for Accounting Review Services [SSARS] No. 1, 1978).

For both reviews, the professional standards suggest that AR and other similar forms of enquiries are the only formal procedures the accountant is required to perform to enable him/her to provide "limited" assurance

regarding the client firm's financial representations. For conventional audits, however, SAS No. 23 requires the auditor to combine the internal control evaluation (unless s/he decides not to rely on AIC) with at least a minimal level of substantive tests. Hence, in situations in which the auditor is to provide such limited assurance, s/he typically relies exclusively on AR to support his/her opinion.

Although the accountants' risk exposure in such engagements is presently unknown, but presumably is less than for actual audits, it still behooves the auditor to employ the most effective analytical review method for such engagements. For example, the information obtained from the performance of such services could be valuable for audit planning decisions.

In addition to its importance in other services being rendered by accountants to their clients, the discussion above indicates that AR can be a cost-effective source of audit evidence. But, to assist the auditor in determining the most effective and efficient combination of audit procedures, research evidence relating to the effectiveness of alternative AR approaches is needed. However, relevant research reported so far has concentrated exclusively on SAR, as the literature review presented in the next chapter indicates.

## CHAPTER 3

### THE ANALYTICAL REVIEW LITERATURE

The goal of enhancing audit effectiveness and efficiency has generated renewed interest in AR because it (AR) is viewed as a reasonably effective and relatively inexpensive source of audit evidence (Holder and Collmer, 1980). Some observers (e.g., Biggs, 1981) have even speculated that the audit of the future is likely to consist only of two elements: control reviews and AR.

Two broad approaches to AR identified earlier are: judgmental analytical review (JAR) and statistical analytical review (SAR). JAR is characterized by the use of insight, experience, knowledge of the client firm's specific environmental data, and professional judgment by the auditor to determine a reasonable range of values for the account balance and to evaluate the significance of any difference with the account book value. SAR, on the other hand, uses formal economic and statistical models to relate the account balance to environmental variables and related account items, as a basis for determining the reasonableness of reported book value.

SAS No. 23 (para. 6) states that the methods selected by the auditor for performing AR are matters for his/her professional judgment. However, GAAS require that effectiveness be the overriding criterion in choosing between audit procedures. Research evidence relating to the relative effectiveness and efficiency of each approach will, therefore, assist the auditor in selecting the most effective AR procedure.

In the following section, I present a review of research reported so far with respect to each AR approach. Thereafter, I will indicate the incompleteness of the existing research, thus forming the basis for a need for the current study.

#### Statistical Analysis Review (SAR)

In order to enhance audit effectiveness and efficiency through the application of AR procedures, some researchers have developed and evaluated the relative effectiveness of various SAR models. For example, Deakin and Granof (1974), Stringer (1975), and Kaplan (1978) have argued in favor of regression analysis in AR. In fact, Stringer (1975) revealed that this method has been a formally accepted AR method by the public accounting firm of Deloitte, Haskins, and Sells (DHS) for a long time. Other writers (e.g., Albrecht and McKeown, 1977; Kinney, 1978) have identified other eligible statistical methods

like the integrated autoregressive-moving-average (ARIMA), Box-Jenkins, Trend Analysis, and Ratio Analysis.

In general, these writers contend that regression analysis works as well as any other complex statistical model based on the criteria of forecast accuracy and/or prediction achievement (Kinney, 1978). Such accuracy and prediction achievement evaluation relate to the degree to which these different statistical methods identify the account book values in which material misstatements are likely to be contained and, hence, are most appropriate for further investigation. As a result, regression-based SAR methods are currently popular in practice.

The primary reason for the suggested application of SAR in (professional) practice is its perceived objectivity, which some observers have assumed would enhance audit effectiveness. For example, Stringer (1975) indicates that

I am convinced that regression analysis provides a more objective basis for performing analytical review, and thereby a means for reasonable quantification of the reliability that may be assigned to this class of substantive tests.

..... Because of the increased objectivity provided by regression analysis, we think it is appropriate to assign a greater portion of the desired reliability from substantive tests to analytical review if regression is used than if it is not (p. 4).



He also notes that another advantage of the regression-based AR is its ability to enhance related reduction in tests of details.

Akresh and Wallace (1980) also claim that the main advantage of the regression model is that it helps the auditor to test the reasonableness, not only of the direction of change, as does the nonstatistical model, but also the amount of change. They declared that "by formalizing the decision process for the acceptance of client representations, the objectivity of the auditor's evidential base is improved" (p. 15).

Incidentally, the preference for SAR seems to be supported by an abundance of research evidence regarding the superior performance of statistical (formal) models over judgmental models reported in the human information processing (HIP) literature (Meehl, 1957; Dawes and Corrigan, 1974; Hogarth, 1980).

Others, however, have focused on a discussion of the methodological and statistical problems inherent in the application of regression-based AR procedures. These include the possible violations of the assumptions underlying regression models and the problem relating to the choice of predictor variables (Warren, 1975; Kinney and Bailey, 1976; Neter, 1980); problems of model validity (Collins, 1981); and the effect of measurement errors on the regression results (Kinney and Salamon,

1979). Some of the likely effects of these factors have been discussed in the accounting literature.

Kinney and Salamon (1979) found that the existence of random measurement errors in the predictor variables can result in biased estimated regression coefficients. This problem leads to a situation in which (1) the account is subject to extensive audit tests when the account is not materially misstated (Type I error), or (2) the account is not extensively tested even though it is materially misstated (Type II error).

Collins (1981) examined empirically the usefulness of the stepwise regression model (SRM) in an auditing context. He indicated that under SRM, variables are often selected on the basis of their ability to effect a reduction in the estimated residual variance, hence there always will exist an understatement of the standard error of prediction. Regarding the effects of errors in the choice of predictor variables, Collins identified the following problems: (i) if irrelevant predictor variables are used, the estimated coefficients will be overstated. In this case, SRM will be oversensitive to unusual fluctuations, thus reducing the efficiency of the AR procedure; (ii) if relevant predictor variables are omitted, the net effect is to make SRM less sensitive to unusual fluctuations, resulting in a reduction in the efficiency of the audit; and (iii) when (i) and (ii)

exist simultaneously, the standard error of prediction will be understated, leading to a reduction in the effectiveness of the audit.

Akresh and Wallace (1980) indicate that, while many of the problems likely to arise from the violation of any assumption often can be corrected, regression analysis might not be applicable to all clients. For example, they suggest that regression analysis should not be used in a time series manner for a client that has had major changes in either the recent base period or the audit period. The same suggestion was made regarding a client that operates in many lines of business. For such clients, they note, regression analysis can be applied to different units of a client firm but not to predicting the book value of an account item.

The amount of data available also may be insufficient for the employment of SAR in case of first-time audits and/or new clients. There will be no reliable (audited) data for the base period in the case of first-time audits, and in the case of new clients, the auditor may not want to rely on audited values reported by the predecessor auditor. Indeed, Wallace (1979) has advised that caution should be exercised in utilizing unaudited data of new clients when conducting SAR. The AICPA's exposure draft on AR (AICPA, 1978, p. 6) also suggests that in such situations, "the auditor should

consider that financial information might not be reliable."

#### Judgmental Analysis Review (JAR)

For reasons other than the problems relating to the application of SAR, as indicated above, some observers have argued in favor of JAR in practice. The main postulated advantage of JAR is that it enables the auditor to employ his/her expertise, experience, knowledge of the client's business operations and industry characteristics, and professional judgment.

Some writers (e.g., Kinney and Felix, 1980) have indicated that JAR is preferable to SAR because statistically-based AR methods necessarily use only a small part of the information which may be available to the auditor. They have, therefore, suggested that research be conducted into the use of expert judgments in AR. This suggestion seems particularly relevant for the following reasons. The accounting literature (e.g., Robertson, 1979. p.343) has acknowledged that AR is essentially judgmental. Yet, Hogarth (1980, p. 4) indicates that no mechanical prediction model can possibly capture the complicated cues, patterns, and other information which humans use for prediction. Hence, as Mock, Biggs, and Watkins (1982) rightly suggest, research directed at understanding the auditor's judgment process

in AR is required before appropriate statistical models can be developed.

Despite the importance and postulated advantages of JAR in the audit process, prior research largely has concentrated exclusively on SAR. The only notable exception is the study by Blocher, Esposito, and Willingham, (1981), in which the authors evaluated certain situational and individual variables on auditor judgment in AR procedures. Consistent with the findings of auditor judgments in internal control studies (e.g., Ashton, 1974; Joyce, 1976; Mock and Watkins, 1980), Blocher, et al found significant variability in auditor judgment for all variables concerning planning, such as the choices which the auditor made in completing the audit program and the resultant time budget. They found, however, that variability of auditor judgment was not significantly affected by the application of analytical review techniques during the audit process.

These findings have important implications for the quality of auditor judgment in the attention-directing and substantive testing aspects of AR. For example, the findings suggest that the planning (attention-directing) phase is likely to be subject to more inconsistent auditor judgments than the usage (substantive testing) phase.

The apparent neglect of auditor judgment in AR seems inappropriate, despite the abundance of evidence

regarding the relative superiority of statistical and formal models over judges noted earlier. Since it has been acknowledged that AR is essentially judgmental, the effectiveness of AR may, therefore, depend ultimately upon the quality of auditor judgments.

In the following section, therefore, I elucidate the need for the study reported herein by describing the nature and importance of auditor judgment in AR. This is followed by a description of the specific research issues which I address.

#### The Need for a Study of Auditor Judgment in AR

The literature review presented above indicates that AR related research reported to date has concentrated exclusively on SAR. This orientation is inappropriate, as the following discussion suggests.

First, some observers have expressed a preference for SAR over JAR because the former is perceived to be more objective than the latter. However, objectivity is neither a necessary nor sufficient condition for choosing among alternative AR methods. Rather, as noted earlier, GAAS indicate that effectiveness is the overriding criterion for choosing between audit procedures or in selecting a combination of audit procedures. Second, the distinction between SAR and JAR noted in the relevant

literature is more apparent than real, because in practice SAR actually requires a combination of judgmental and statistical methods. In practice, SAR involves the auditor (a) identifying the relevant predictor variables, (b) determining the reliability levels and precision limits, (c) evaluating the plausibility of relationships suggested by the statistical analysis, and (d) subjectively deciding what action to take. The role assigned to the statistical model is merely to weight, perhaps optimally, the variables suggested by the auditor. In other words, the statistical aspect of SAR merely supplements human (auditor) knowledge and skill in practice. Given the pervasive role of auditor-judgment in AR tasks, it is inappropriate not only to assume the separate existence of a purely statistical AR approach, but also to suggest that the (nonexisting) "statistical" AR is preferable to JAR.

Third, many public accounting (CPA) firms rely only on JAR, while others employ a combination of statistical and judgmental methods commonly known as SAR. Therefore, it is inappropriate to concentrate to concentrate research efforts only on the statistical aspect of AR. Fourth, even if a purely statistical AR method were to exist, there are situations in which it would not be applicable in practice. Discussions with practicing auditors indicate that SAR is applied only to

large client firms, mainly because of the high cost involved in setting up SAR for each client firm. Therefore, JAR still is the conventional method for many clients even by those CPA firms which have an inclination towards enhancing the objectivity of the AR procedure. Also, as indicated earlier, the amount of data available may be insufficient for the employment of SAR in the case of first-time audits and/or new clients.

Furthermore, Akresh and Wallace (1980) indicate that, while many of the problems likely to arise from the violation of any assumption often can be corrected, the statistical method might not be applicable to all clients. For example, they suggest that regression analysis should not be used in a time series manner for a client that has had major changes either in the recent base period or the audit period. The same suggestion was made regarding a client that operates in many lines of business. For such clients, therefore, regression analysis can be applied to different units of a client firm but not to predicting the book value of an account item. This observation suggests an additional limitation on the use of SAR-based ARP in practice.

Finally, Kinney (1979) acknowledged that there are various specific local and other internal-to-the-firm data which are not available from published sources, which enable expert auditors to be quite skilled at predicting



potential audit adjustments. Furthermore, some of the evidence gathering procedures available to the auditor are not amenable to statistical modeling. For example, much of the information which influences audit judgment is derived from verbal discussions and written representations of client personnel and independent parties.

Given that JAR will continue to play a significant role in practice, it is desirable to know auditors' ability to detect account items which may require adjustments, as well as the factors which might affect these judgments. This desire is underscored by the fact that the ultimate effectiveness and efficiency of the audit may depend principally on auditor judgments in PAR tasks, since the initial allocation of audit efforts are made at this stage of the audit. That is, given the importance of the accuracy of auditor judgment in PAR on the allocation of audit efforts, it is desirable to have evidence relating to how much the auditor knows, how much s/he knows that s/he knows, and the factors that may influence his/her judgments. Also, to enhance an understanding of the auditors' decision process, it is desirable to identify the types of information they consider relevant for such (PAR) tasks. The following research issues addressed in this study will provide evidence bearing on these matters.

### Research Issues: An Outline

To provide a rationale for the research issues addressed in this study, I summarize the substance of the discussion above as follows.

First, the discussion indicates that AR is an important tool for the audit planning process. Second, it indicates that AR can be an effective but relatively inexpensive source of audit evidence. Third, there are situations in which SAR may not be applicable. In addition, there are numerous methodological and statistical problems inherent in the application of SAR, as discussed earlier. Finally, the discussion suggests that AR is ultimately a judgmental task, even in situations in which statistical models are used to provide evidence as inputs into the auditor's judgmental process. This view is buttressed by an acknowledgment in the relevant literature (e.g., Robertson, 1981, p. 367) that the AR (and other supplementary) procedures are highly judgmental. Yet, no research has so far provided direct evidence regarding the accuracy of auditor judgments in PAR.

The aim of this dissertation is to fill this void, by providing evidence that will enable one to appraise the implications of several aspects of auditor judgments for the effectiveness and, ultimately, the efficiency of the audit. To achieve this objective, I provide evidence

bearing on the following research issues, the details of which are discussed fully in Chapter 5.

Research Issue Number One:

Detectability of Auditor Judgments in PAR

As indicated earlier, the main objective of PAR is to assist the auditor in identifying the account book values which may contain material errors. The greater the accuracy of auditor judgments, the more the effectiveness and efficiency of the audit is enhanced, since s/he will then be able to concentrate audit efforts in the areas likely to contain material misstatements. This study, therefore, provides evidence bearing on the degree of accuracy of auditor judgments in PAR.

Research Issue Number Two:

Implicit Loss Functions  
Affecting Auditors' PAR Judgments

Many research studies have reported that human beings are generally suboptimal decision makers under uncertain conditions. They indicate that, in most cases, normative models tend to outperform human judges. Some of the explanations offered for these findings include the possibility that human decision makers are not able to optimally process the information provided in experimental tasks, and that they have limited memory, which limits the

amount of information they can process at a time. Consequently, the arguments continue, their judgments are less accurate than those of normative models.

An important variable which these studies have consistently failed to address is the effect of implicit loss functions on decision makers' judgments in experimental tasks. If the decision makers employ loss functions which differ from those assumed by normative models, then no meaningful comparison could be made between the accuracy of the normative model and the accuracy of the decision maker.

What is required, therefore, is evidence regarding the implicit loss function(s) which decision makers employ in an experimental task as a basis for explaining performance. This evidence is obtainable from an evaluation of the decision rules employed by the subjects. Such evidence is particularly important in the context of this study, since the decision rules employed could have impact upon the effectiveness and efficiency of the audit. Hence, in this study, I provide evidence relating to the decision rules and, consequently, the implicit loss functions, employed by the subjects in the experiment, and discuss their implications for audit effectiveness and efficiency.

### Research Issue Three:

#### Effect of Nature of Task on Auditors' PAR Judgments

One can make a distinction between experimental tasks which request for subjects' perceptions under uncertain conditions, and those which require the subjects to specify what actions they will take under uncertain conditions. The first type merely asks for feelings regarding the items or content of an experiment, ordinarily in terms of subjective probability judgments. The second type requests the subjects to make actual decisions based on their assessment of the information provided. Howell and Burnett (1978) have classified these as prediction and choice (action) tasks, respectively.

However, a major difference between the characteristics of the two types of tasks which most earlier studies have not considered is the effect of the perceived importance of the task in each of these situations on the subjects' judgments. For example, in the first situation, the subjects normally will not be concerned with the consequences of their judgments, while in the second, the subjects should be more concerned about the consequences of their decisions (see Tukey, 1960; Howell and Burnett, 1978).

In this experiment, the subjects were asked to provide responses under both conditions. The former

provides a measure of the subjects' perception of the accuracy of the stated book values. The latter provides an approximate measure of what the subjects would have done in practice and, hence, is of direct interest to this study. A comparison of their decision rules under both situations also enables one to determine the extent of correspondence between the subjects' beliefs and their preferences.

#### Research Issue Number Four:

##### Effect of State of AIC on Auditors' PAR Judgments

As indicated earlier, the auditor relies on the AIC to reduce the risk that material errors will occur in the accounting process used to develop the financial statements. The level of reliance which the auditor places on an AIC is positively related to its perceived strength, since s/he expects a lower probability of material error occurrences when the AIC is perceived to be strong. Similarly, s/he expects a higher probability of occurrence of material errors when the AIC is adjudged weak. The auditor will be inclined to perform fewer tests of details in the first case than in the second.

Specifically, if the AIC is adjudged strong, the auditor will be inclined to flag fewer account items for intensive audit relative to the number of account items

s/he would have flagged for intensive audit had the AIC perceived to be weak. This study provides evidence relating to this idea.

Research Issue Number Five:

Effect of Functional Level  
on Auditors' PAR Judgments

The expected relationship between functional level and quality of job-related judgments under uncertain conditions is predicated on the following.

First, the relevant literature suggests that performance in job-related tasks may depend upon the amount of substantive knowledge possessed by the subjects. In particular, some researchers (e.g., Charness, 1976; Chase and Simon, 1973) have indicated that, through repetitive performance of job-related tasks, decision-makers accumulate relevant information about the intricacies of the task. They indicate that the development of expertise depends, to a great extent, on storage in long-term memory of a series of meaningful cue patterns which are prototypical of certain class memberships.

Second, Waller and Felix (1982) have discussed the process through which auditors accumulate job-related experience, and how such experience enhances auditors' performance. They indicate that the auditor must acquire

knowledge regarding, among other matters, (a) professional standards for performing an audit and reporting the results thereof; (b) how to plan and execute the process of evidence collection and evaluation; and (c) the types, qualities, and interaction of evidence with the client environment. Some of this knowledge, the authors indicate, may be acquired through formal instruction. However, they admit that, for the most part, the auditor's cognitive structures that represent his/her knowledge of the practice of auditing, and which drive his/her perception and judgments, are the product of experiential action and observation. Furthermore, Waller and Felix suggest that the professional auditor's internal representation of the opinion formulation process is likely to develop as a hierarchical network of interactive, declarative, and procedural knowledge structures which are built upon experiential data. These data, they indicate, include observations of event co-occurrences and action-outcome feedback co-occurrences.

Third, the accounting literature indicates that good performance in PAR tasks requires a familiarity with, and an understanding of, the nature of the client firm's industry, business operations, accounting procedures, as well as other qualitative factors like the perceived quality of personnel and the integrity of management. The longer an auditor is associated with a client firm, the



better his/her understanding of these factors. The discussion above suggests that the more experienced auditors should have a greater understanding of the nature of the task because of relevant knowledge accumulated over time. As a result, they should outperform the less experienced auditors in job-related tasks. Specifically, I hypothesize that audit managers' judgmental accuracy should be higher than those of the audit seniors. Similarly, I hypothesize that managers' decision errors should be less than those of seniors. This study provides evidence bearing on this idea.

#### Research Issue Number Six:

##### Auditors' Sensitivity to Their Degree of Uncertainty

When making judgments under uncertain conditions, a subject may be oversensitive or undersensitive to his/her degree of uncertainty. In such situations, the subject's probabilistic judgments may not conform with the stochastic process underlying the events of interest. The relevant literature (e.g., Beck, Solomon, and Tommasini, 1982) has indicated that such nonconformance may lead to audit effectiveness and efficiency errors.

This study, therefore, provides evidence regarding the auditors' degree of sensitivity to their uncertainty in PAR tasks.

**Research Issue Number Seven:**

**Information Items Required by Auditors  
to Facilitate Their PAR Judgments**

The accounting literature (e.g., Abdel-Khalik and El-Shesai, 1980) has indicated that the type of information chosen by subjects may affect the accuracy of their judgments in experimental tasks. Also, in recognition of the potential effects of differences in information search and choice behavior on the degree of judgment consensus among auditors, some researchers (e.g., Mock, et al, 1982) have suggested that studies of the information which auditors use for AR judgments be performed. So far, no study has been reported which provides direct evidence bearing on this issue.

In this study, therefore, I provide evidence bearing on the relative importance of information items which auditors consider relevant to facilitate their PAR judgments. To enhance the validity of the evidence provided on these research issues, it is necessary to apply statistical measures that are appropriate for this purpose. For example, two prior studies (Shuford and Brown, 1975; Lichtenstein and Fischhoff, 1977) have dealt explicitly with the problem of knowing that one knows, and have attempted to measure it in relation to the extent of knowing. Both studies make use of the concept of

calibration of subjective probability estimates, an approach to which other researchers (e.g. Ferrell and McGoey, 1980; Hosseini-Ardehali, 1981) have related some theoretical objections.

To provide justification for the use of a statistical model based on SDT for analyzing this study's data, I present below a discussion of the concept of calibration and a summary of its previously identified limitations as a measure of knowing that one knows. Finally, I present an argument regarding the need to provide evidence bearing on the nature of calibration of the responses of this study's subjects.

#### Calibration of Subjective Probabilities

Calibration is concerned with the appropriateness of assessors' confidence in their subjective judgments. It has variously been called Realism of Confidence (Adams and Adams, 1961); Appropriateness of Confidence (Oskamp, 1962); Secondary Validity (Murphy and Winkler, 1971); Realism (Brown and Shuford, 1973); Reliability (Murphy, 1973); and External Validity (Brown and Shuford, 1973).

Calibration measures the correspondence between the level of confidence an assessor has in his/her probabilistic judgments and the proportion of times those judgments are true. For example, over the long run, for

all judgments assigned a probability of .63, 63% should be true if the assessor is to be adjudged well calibrated. When an entire probability distribution is assessed, calibration refers to the correspondence between the fractiles of an ensemble of prior probability distributions (PPDs) and the relative frequency with which the actual outcome of the uncertain event falls at or below the specified fractile values. For example, 75% of, say, audit values should fall at or below the values assessed for the .75 fractiles of an ensemble of well calibrated PPDs (see Lichtenstein, Fischhoff, and Phillips, 1982). Calibration may be reported using calibration curves, as shown in Figure 3-1.

Calibration curves can be derived through the following steps described by Lichtenstein, et al, (1982):

- (a) collect answers and subjective probabilities of answers to a set of items whose "true" values are, or will shortly be, known to the researcher;
- (b) categorize the subjective probabilities,  $r$ , if they were not restricted in the first place. For example, all responses between .60 and .69 are placed in the same category, say  $r = .65$ ;
- (c) compute for each  $r$  category the proportion  $c$  (i.e.,  $P(c/r)$  of correct responses, and
- (d) for each category, plot the proportion correct against the nominal (assessed) probability.

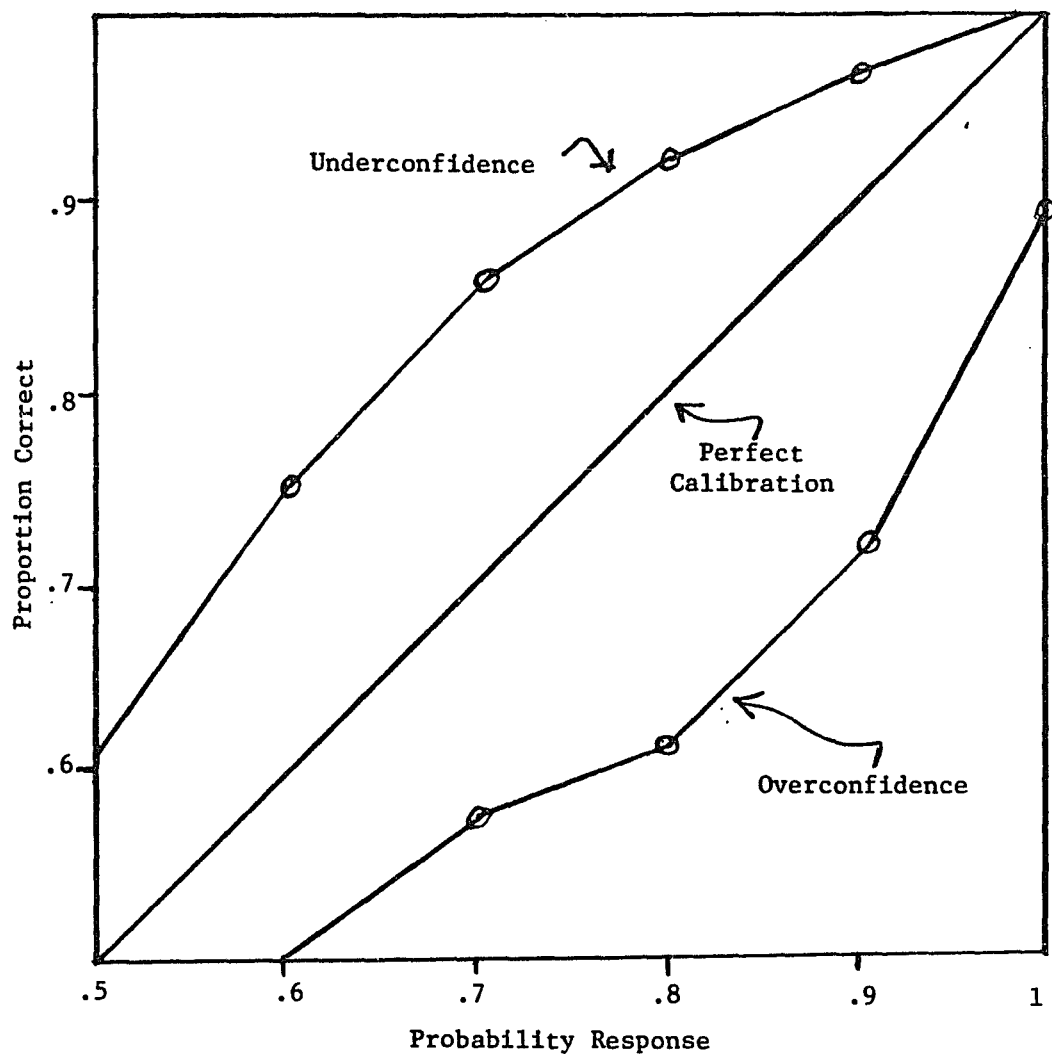


Fig. 3-1. Example of Calibration Curves for a Two-Alternative, Forced-Choice Task

Calibration curves may indicate underconfidence or overconfidence. In case of the former, the proportion correct is greater than the probability assigned; for the latter, the probability assigned is larger than the proportion correct. Figure 3-1 illustrates each of these situations.

Several measures of overall calibration have been suggested. The overall tendency for a judge to be overconfident or underconfident is measured by

$$\text{Over/underconfidence} = 1/N \sum_{t=1}^T n_t (r_t - c_t) \quad (3-1)$$

where  $N$  is the total number of responses,  $n_t$  is the number of times the response  $r_t$  was used,  $c_t$  is the proportion of all items assigned probability  $r_t$ , and  $T$  is the total number of different response categories. Overconfidence is shown by a positive difference, underconfidence by a negative difference.

A measure of the adequacy of calibration proposed by Oskamp (1962) replaces the parenthesis in equation (3-1) by an absolute value sign, thereby measuring the mean weighted distance between the calibration curve and the identity line. An alternative measure proposed by Murphy (1973) is to take squared deviations:

$$\text{Calibration} = 1/N \sum_{t=1}^T n_t (r_t - c_t) \quad (3-2)$$

A perfectly calibrated person would score 0 on this measure. The worst possible score, 1.0, can be obtained only by a respondent who always responds  $r = 1.0$  when wrong and  $r = 0.0$  when right.

Murphy (1973) has shown the calibration score in equation (3-2) to be an additive component of a widely used quadratic scoring rule (i.e., Brier Score) for probability assessments. The Brier Score (Brier, 1950) for the discrete case is defined as

$$S_B = 1/N \sum_{k=1}^K (r_k - c_k)^2 \quad (3-3)$$

where  $K$  indexes the  $N$  events for which a probability was assessed and  $c = 1$  or  $0$  depending on whether the event occurred or not. It is assumed that the probability of only one of the mutually exclusive and exhaustive events is assessed. The Brier Score can be further partitioned as

$$S = P(C)[1-P(C)] + 1/N \sum_{t=1}^P n_t (r_t - c_t)^2 + 1/N \sum_{t=1}^P n_t [c_t - P(C)]^2 \quad (3-4)$$

where  $t$  indexes the response categories which have  $n$  responses,  $p(c)$  is the overall proportion correct, and  $c$  is the proportion correct in each response category. The first term measures knowledge, the second calibration, and the third term measures resolution. Resolution reflects

the degree to which assessors can discriminate among different degrees of uncertainty, independent of numerical labels assigned. The best Brier Score is a minimum of 0.0, and the worst a maximum of 1.0 attainable when all answers are wrong and all are assigned 1.0.

The discussion above indicates that calibration essentially measures the extent of accurate quantitative representation of one's uncertainty. In the next section I review the literature on knowing that one knows, and discuss the limitations of measuring this attribute using the concept of calibration. Thereafter, I present an argument to the effect that calibration measures are of direct relevance to the objectives of this study, regardless of its limitations as a measure of knowing that one knows.

#### Prior Research on Knowing That One Knows

Shuford and Brown (1975) addressed the question of knowing that one knows in an educational setting. They stated that a student's choice of an answer to a multiple-choice test question is a coarse measure of his/her knowledge about the subject matter of the question. Much finer measurement, they suggest, might be achieved if the student were asked to estimate, for each possible answer, the probability that it is the correct one. Shuford and Brown also assume that each student



wants to maximize his/her score, the measurement of which they said requires a scoring rule that possesses the property that a student can maximize his/her expected score if and only if his/her probability response matches his/her true feelings about the likelihood of correctness of the answer. The authors indicate that the logarithmic scoring rule probably satisfies this criterion better than any other scoring rule.

Using this scoring rule, Shuford and Brown (1975) calculated what they interpreted as an index of the amount of information each student perceives s/he possesses with respect to the subject matter (say,  $I_p$ ). They then derived a measure of the information the student actually possesses (say,  $I_a$ ) with respect to the test's subject matter. The authors then define  $I_a$  as a measure of knowledge, and  $|I_a - I_p|$  as a measure of not knowing that one knows (i.e., the discrepancy between perceived and actual knowledge).

However, Hosseini-Ardehali (1981) has expressed concern about the validity of these measures. First, she said that it is not clear how Shuford and Brown (1975) think knowledge and knowing that one knows should be measured, since at one point they identify it with good calibration; that is, "the ability to assess their uncertainties accurately". Second, she indicated that the authors use an information measure to assess these

attributes which, she argued, also depends upon the degree of calibration. She then declares that "in any case, good calibration is at the heart of their conception of knowing that one knows" (p. 26).

Given this premise, Hosseini-Ardehali (1981) raised two kinds of objections regarding Shuford and Brown's (1975) measures of knowledge and knowing that one knows. The first technical but minor objection relates to the use of a linear fit to the calibration by the authors. The second and major objection is to equating knowing that one knows with accurate quantitative representation of one's uncertainty. Hosseini-Ardehali emphasized that knowing that one knows tends to give extreme probabilities and knowing that one knows the answers are conceptually different. She said, in Shuford and Brown's words, the ability "to discriminate with great accuracy what they know well from what they know less well" and ability to "assess their uncertainties accurately".

Lichtenstein and Fischhoff (1977) attempted to provide an answer to the question, "Do Those Who Know More Also Know More About How Much They Know". The criterion they used for measuring knowing that one knows was degree of calibration. The main findings of their study are:

- (a) subjective probability judgments are mostly overconfident (i.e., more extreme than the corresponding relative frequency;

- (b) there is a systematic trend from overconfidence toward underconfidence as the overall proportion of correct responses increase;
- (c) there is no difference between the calibration of those who are more expert or intelligent and those who are less so when the overall proportion of correct responses is the same; and
- (d) calibration can improve with training.

Based on these findings, Lichtenstein and Fischhoff conclude that those who know more do not know more about how much they know. However, Hosseini-Ardehali (1981) indicates that, since these authors also use calibration as their criterion, the objections raised against Shuford and Brown (1975) apply to them. She asserted that calibration per se is not a measure of knowing that one knows, but just the quality of its numerical expression.

Nevertheless, evidence regarding the calibration of the responses of this study's subjects will be provided for the following reasons. First, as will be shown in the next chapter, this study does not satisfy the condition specified by Hosseini-Ardehali (1981) which enables one to measure knowing (i.e., knowledge) independent of knowing that one knows. Second, the accounting literature (e.g.,

Beck, et al, 1982) indicates that the nature of (mis)calibration of auditor judgments has implications for audit effectiveness and efficiency errors, since overconfident auditors are likely to collect less than adequate audit evidence as a basis for their opinion. Evidence relating to the calibration of the subjects' responses is, therefore, of direct relevance to the objectives of this study and, hence, was used to evaluate the auditors' sensitivity to their level of uncertainty.

## CHAPTER 4

### SIGNAL DETECTION ANALYSIS

As indicated in Chapter 2, both the importance of AR in the audit process and the pervasive role of the auditor judgment in AR suggest the need for evidence relating to (a) how much the auditor knows, (b) how much s/he knows that s/he knows, and (c) the decision rules employed for making his/her judgments.

Prior research has employed calibration of subjective judgments as a measure of knowing that one knows. However, some researchers (e.g., Ferrell and McGoey, 1980) recently have questioned the adequacy of calibration as a measure of knowing that one knows. They, instead, have suggested that a model based on signal detection theory, called the Decision Variable Partition Model (DVPM), is more appropriate than calibration for measuring knowing that one knows.

In this chapter, I present a discussion of (a) signal detection tasks, followed by a discussion of (b) signal detection theory. I also discuss DVPM based on the signal detection theory, which provides a measure of (a)

how much the subject knows, (b) knowing that one knows, (c) the decision rules employed by the subjects in a signal detection task, and (d) calibration of subjective judgments. Thereafter, I discuss the limitations of calibration as a measure of knowing that one knows. Finally, I argue that PAR tasks are analogous to signal detection tasks, to justify the use of a model based on signal detection theory for providing evidence bearing on the research issues identified earlier.

#### Signal Detection Tasks

The task of the detector (i.e., observer or subject) is to decide, on the basis of noisy (uncertain) evidence, whether the observed stimulus resulted from one category, usually called signal plus noise (SN), or from the other category, called noise (N). The fundamental detection task, therefore, involves the following two principal features:

- (a) the subject observes data in which there are random or probabilistic occurrences of two events, SN and N;
- (b) after each observation, the subject makes a decision "yes", (sn), corresponding to SN, or "no", (n), corresponding to N.

These decision/state combinations have four possible outcomes: (1) [SN,sn], i.e., a hit; (2) [SN,n], a miss;

(3)  $[N,sn]$ , a false alarm, and (4)  $[N,n]$ , correct rejection (see Figure 4-1).

The measures of performance in SDT use only the hit rate  $p(sn/SN)$  (i.e. the conditional probability of a "yes" response given that the signal occurred), and the false alarm rate  $p(sn/N)$  (i.e., the conditional probability of the "yes" response given that the signal did not occur). The hit and false alarm rates can be computed from the frequencies ( $f$ ) of the occurrence of the observable events ( $SN,N$ ) and the two observable responses ( $sn,n$ ). This computation of hit and false alarm rates is illustrated in Figure 4-1, in which  $f_1$  refers to the number of hits,  $f_2$  the number of misses,  $f_3$  the number of false alarms, and  $f_4$  the number of correct rejections.

#### Signal Detection Theory

The fundamental idea underlying signal detection theory (SDT) is that detection of a signal or pattern under noisy (uncertain) conditions involves three distinct but related steps: (a) observing data, (b) organizing or processing the data, and (c) making a decision. SDT is applicable to those situations in which two (or more) classes of events are to be discriminated, on the basis of evidence which does not unequivocally support one of a number of hypotheses.

		R E S P O N S E	
		sn	n
S T I M U L U S	SN	Hit	Miss
	N	False Alarm	Correct Rejection

Where Hit and False Alarm Rates Are Determined as Follows:

		R E S P O N S E		
		sn	n	
S T I M U L U S	SN	$f_1$	$f_2$	$P(\text{sn}/\text{SN}) = f_1 / (f_1 + f_2)$
	N	$f_3$	$f_4$	$P(\text{sn}/\text{N}) = f_3 / (f_3 + f_4)$

Fig. 4-1. Example of a Decision-Making Matrix



The basic assumption underlying SDT is that each decision is made as if the observer, upon observing an event (or decision variable), say  $X$ , uses a statistic derived from the many characteristics of that event (Pastore and Scheizer, 1974). The optimum statistic for the decision is the likelihood ratio  $L(x) = p(x/SN)/p(x/N)$ . It is the relative likelihood that the observation  $x$  came from one as opposed to the other class of events. SDT also assumes that the subject decides on a cut-off value of  $L(x)$ , say  $c$ , and that the decision relating to any  $x$  is simply a statement of whether  $L(x) > c$ . The value of  $c$  is assumed to depend on  $p(SN)$  and both the rewards of correct, and costs of incorrect, responses (Sheridan and Ferrell, 1981).

The ability of the subject to discriminate between the two classes of events ( $SN$  and  $N$ ) is inversely related to the total area common to the two conditional probability density functions  $f(x)$ , ( $i = 1, 2$ ). The hit rate,  $p(sn/SN)$ , will be the integral of  $f(x/SN)$ , the probability density function (pdf) of  $x$  conditional on  $SN$ , over the portion of the axis for which  $x > c$ . Likewise, the false alarm rate,  $p(sn/N)$ , will be the integral of  $f(x/N)$  over the same region. This relationship is illustrated in Figure 4-2.

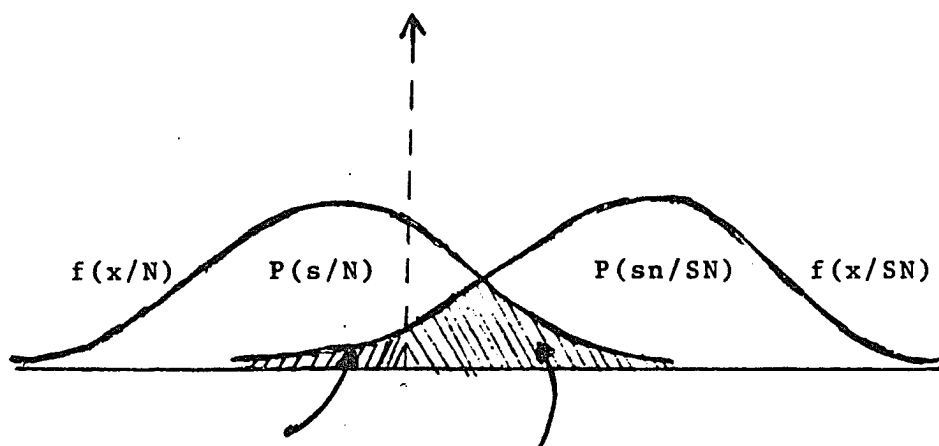


Fig. 4-2. Underlying Distribution for SN, N.

### Relative Operating Characteristics

The relative operating characteristic (ROC) is the locus of points representing the performance of a subject across all judgment criteria under a fixed experimental condition (Pastore and Scheirer, 1974). It is a curve whose overall location corresponds to a particular degree of discrimination while the position of any point along the curve represents a particular degree of bias (judgment criterion) in the observer's response (see Swets, 1973, p. 991).

The ROC curve essentially represents the subject's ability to discriminate between the two events SN and N, each point on the curve representing a pair of hit and false alarm rates. It is used, graphically, to specify the relation between hit and false alarm rates. The ordinate of the ROC function is  $p(\text{sn}/\text{SN})$ , and the abscissa is  $p(\text{sn}/\text{N})$ , for each judgment criterion. Its shape is affected by, among other things, the extent of overlap of the underlying density functions of the events of interest. A typical ROC curve is illustrated in Figure 4-3.

The effect of the extent of overlap of the SN and N events' distributions on the shape of the ROC curve can be illustrated with the following example. Suppose that the two conditional distributions of the decision variable X differ very little. This means that the false alarm

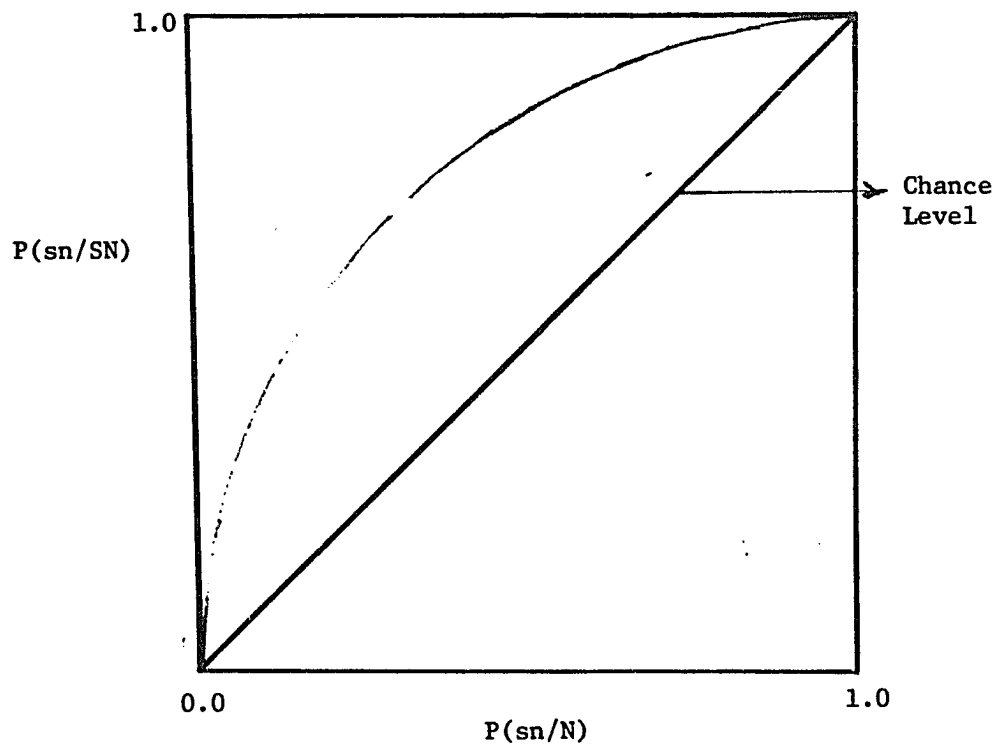


Fig. 4-3. The ROC Graph

rate and the hit rate will be almost equal regardless of the cutoff value  $c$ . Detection performance in this case is likely to be near chance level, represented by the diagonal line in Figure 4-3. If the two distributions differ considerably or, equivalently, have a low degree of overlap, the hit rate may be much greater than the false alarm rate given an appropriate cutoff, and the observer's detection performance could be quite high.

When the decision variable  $X$  is monotonically increasing with likelihood ratio  $L(x)$  associated with the cutoff value  $c$  for that point, the slope of a segment of the ROC curve joining two adjacent points is numerically the same as the likelihood ratio. This can be illustrated by differentiating the expressions for the hit and false alarm rates with respect to the decision variable  $X$  as follows:

$$\begin{aligned} dP(s_n/SN)/dx &= -f(x/SN) \\ dP(s_n/N)/dx &= -f(x/N) \end{aligned} \tag{4-1}$$

from which, with the use of chain rule, one derives

$$[dP(s_n/SN)/dP(s_n/N)] = -f(x/SN)/[-f(x/N)] = L(x) \tag{4-2}$$

which is identical to  $f(Z/SN)/f(Z/N) = L(Z)$ . When  $X$  is not monotonically increasing with  $L(x)$ , the slope of the ROC curve is the likelihood ratio based on the decision variable  $X$  itself.

When the underlying distributions are normal and have equal variance, the ROC curve has a monotonically

decreasing slope. With unequal variances, two normal distributions will intersect at two places, and the decision performance cannot be optimal with a single cutoff.

#### Indices of Detectability

An observer's detectability which, as indicated earlier, measures his/her ability to discriminate between two events SN and N, is related to the extent of overlap of the two conditional distributions. Several measures of detectability have been proposed, each of which is dependent on the assumed underlying distributions for the events SN and N.

If one assumes that the two distributions are Gaussian normal with equal variances, their separation can be described by a single parameter  $d'$ :

$$d' = (U_{sn} - U_n)/r \quad (4-3)$$

where  $u$  is the mean and  $r$  the (common) standard deviation. Given experimental data, the normal distribution can be fitted to it and  $d'$  calculated. The probabilities  $P(sn/SN)$  and  $P(sn/N)$  then are transformed to  $Z$  coordinates,  $Z(sn/SN)$  and  $Z(sn/N)$ , according to the relation

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z \exp \left\{ \frac{-x^2}{2} \right\} dx \quad (4-4)$$

i.e., the integral of the standard normal distribution. A straight line of unit slope is fitted to the points on Z coordinates. The index d is the value of  $Z(s_n/N)$  at which  $Z(s_n/SN) = 0.0$

But when the variances are not equal,  $d'$  cannot be used. Instead, two other measures are commonly adopted. The first,  $m$ , is the distance between the means of the signal and noise distributions measured in standard deviation units of the noise distribution. It is equal to  $z(S/N)$  at the point on the ROC curve where  $z(s/SN)$  is equal to zero. If  $U_s$  is the SN's distribution mean, and  $U_n$  is the mean of the N distribution, one can define

$$m = (U_s - U_n)/\sigma_n \quad (4-5)$$

and, from the definition given earlier, it follows that

$$m = z(s/N) - z(s/SN) \quad (4-6)$$

at the point where  $z(s/SN)$  is equal to zero. The virtue of  $m$  is that it is directly related to the distance between the distributions' means.

A second method for measuring detectability when the variances of the underlying distributions are unequal, proposed by Egan and Clarke (1966), is  $d'e$ . It is defined as twice the absolute value of  $z(s/SN)$  or  $z(s/N)$  at the point where the ROC curve intersects the negative diagonal. The convention here is to read the value of  $z(s/SN)$  or  $z(s/N)$  at the point where the ROC curve intersects the negative diagonal.

Many reasons have been advanced for using  $d'e$ . First, it gives equal weight to units of both the SN and N distributions, whereas in the case of  $m$ , detectability is scaled in units of the N distribution alone. A second argument which makes  $d'e$  a desirable measure of detectability is that the point from which it is read from the ROC curve normally falls within the range of points produced by the observers' criteria. In the case of  $m$ , the criteria will not give points which lie as far down the ROC curve as the point from which  $m$  is read. Consequently, the  $m$  point may have to be obtained through extrapolation, with the attendant errors associated with such exercise. Third, Egan and Clarke (1966) report that the slopes of ROC curves tend to be unstable and to vary somewhat from session to session for the same observer. These changes in slopes, they indicate, appear to alter the value of  $m$  more than the value of  $d'e$ , thereby making the latter a more stable measure.

Nevertheless, Green and Swets (1966) indicate that it does not matter which measure one calculates. They suggest that it is possible to derive the values of  $m$  or  $d'e$  through the following conversion formula:

$$d'e = 2 m(s/[1+s]) \quad (4-7)$$

where  $s = [dz(sn/SN)]/[dz(sn/N)]$ ; that is,  $s$  is the slope of the ROC curve. Similarly, it can be shown that

$$m = d'e(1+s)/2s \quad (4-8)$$



Another measure of detectability suggested by Green (1964) is the area under the ROC curve which, as indicated earlier, is the relation between  $P(sn/SN)$  [i.e., the hit rate] and  $P(sn/N)$  [i.e., the false alarm rate]. A typical area under the ROC curve is illustrated in Figure 4-4.

#### Experimental Approaches in SDT

Three modes of eliciting responses typically are used in SDT studies: (a) yes-no, (b) two-alternative, forced-choice (2AFC), and (c) rating scale response modes.

##### (a) "Yes-No" Mode

The subject is presented with sample data from SN and/or N and asked for a response sn or n. It is assumed, as in other experimental methods, that the subject is rewarded for correct responses according to a well-defined pay-off matrix. It also is assumed that the subject responds "yes" (sn) only if his/her decision variable is greater than some cut-off value  $c$ , and "no" (n) otherwise. In SDT, the appropriate decision variable is the likelihood ratio  $P(SN)/P(N)$  because it is assumed that there is a strictly monotonic relation between  $L(x)$  and  $P(sn/x)$  (Egan, 1975, p. 19). This method does not account for differences in confidence in the responses; therefore, all that can be said about the subject's

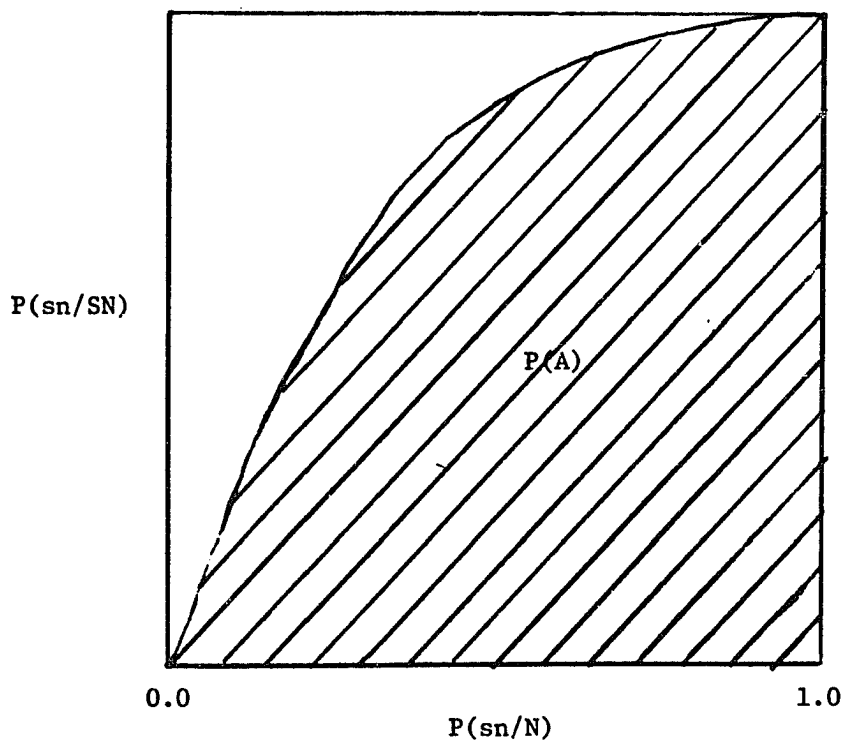


Fig. 4-4. ROC Curve as a Measure of Detectability

decision variable for a given stimulus is that it is greater (or less) than his cut-off value.

(b) Two-Alternative, Forced-Choice Mode

The subject is presented with instances consisting of two stimuli, one SN and the other N, in random order. The subject then indicates the one s/he believes to be SN (Sheridan and Ferrell, 1981). Optimal performance would be to respond "sn" to the interval producing the higher likelihood ratio. If this strategy is employed, the probability of a correct response on a 2AFC task should be equal to the area under the ROC curve from "yes-no" and rating scale experiments.

Cognitively, this task should be relatively easy for the observer, since one sample from each pair must come from each population, and s/he simply makes a comparison judgment as to which has the larger (smaller) value in relation to his/her decision variable (that is, the observer's subjectively determined cutoff criterion for responding "sn" or "n"). The 2AFC format also is particularly useful, since its score, say,  $p(c)$ , provides a nonparametric measure of a subject's detectability, independent of the nature of the underlying conditional distributions (Sheridan and Ferrell, 1981).

### c) Rating Scale Mode

Rating scale and yes-no experiments are identical, except that in the former case it is assumed that the subject can respond with an indicator of the relative value of the scalar decision variable that results from an observation. This method makes it possible to generate an approximation of the underlying distributions, and to infer from them the ROC curve resulting from decisions based on the use of a cut-off  $c$  on that variable.

For example, one can present a subject with instances of SN or N, and ask him/her to respond with a subjective posterior probability that the stimulus is SN or N. In addition, s/he is asked to express his/her level of "certainty" on a scale of, say, 1 to 5, from which the distribution of responses  $r$  can then be constructed. Each  $r$  must be unambiguously defined for the subject. For example, each  $r$  on a 1 to 5 rating scale can be defined as follows:

- 1 = "quite certain it was N",
- 2 = "fairly certain it was N",
- 3 = "as likely to be N as SN",
- 4 = "fairly certain it was SN",
- 5 = "quite certain it was SN".

If one assumes that decisions are to be made using a single cutoff, an approximate ROC curve can be generated by taking each response category as a cutoff and

calculating the hit and false alarm rates for it, using the relative frequencies as approximations to the conditional probabilities. The hit and false alarm rates are then calculated as follows:

$$\begin{aligned} P(\text{sn}/\text{SN})_k &= \sum_{i=1}^5 P(i/\text{SN}) \\ P(\text{sn}/N)_k &= \sum_{i=1}^5 P(i/N), \text{ where} \end{aligned} \quad (4-9)$$

$$P(i/\text{SN}) = \frac{\text{The number of responses "i" when SN is the case}}{\text{The number of times SN was the case}}$$

$$P(i/N) = \frac{\text{The number of responses "i" when N is the case}}{\text{The number of times N was the case}}$$

The number of useful points from a rating scale task for the ROC will always be one less than the number of rating categories used (McNicol, 1972, p. 28). If there are five response categories, there are four cutoff points:  $x_1, x_2, x_3, x_4$ , such that the observer will respond: 1 if  $x < x_1$ ; 2 if  $x_1 < x < x_2$ ; 3 if  $x_2 < x < x_3$ ; 4 if  $x_3 < x < x_4$ ; and 5 if  $4 < x$ . This method yields a point on the ROC curve for each cutoff point.

The relevant literature has identified several advantages of the rating scale mode, some of which are discussed as follows. From the ratings provided by the subjects, it is possible to determine the related ROC curves. The area under such an ROC curve, say  $A_c$ , has

been proposed by Green (1964) as a nonparametric measure of detectability. The importance of this measure cannot be overemphasized because, as McNicol (1972) notes, in a detection task one does not directly observe underlying SN or N distributions. Rather, he said, the data one has consists of sets of hit and false alarm rates from which the underlying distributions must be inferred. The rating method also makes use of the fact that during a single series of observation intervals, the human observer is capable of adopting multiple decision rules (Egan, Schulman, and Greenberg, 1959). Pastore and Scheirer (1974) note, therefore, that such a nonparametric model of SDT whose measures are independent of the exact nature of underlying distributions would be of general utility.

The rating scale mode also compares favorably with the other methods on other dimensions. Green (1964) reported that the area under the ROC curve obtained with the Yes-No mode is equal to the percentage of correct responses in the two-alternative forced choice mode. His most important finding was that no assumptions had to be made about the form of the conditional distributions or the decision variable  $X$ . Furthermore, Egan, Schulman, and Greenberg (1959) indicate that the rating mode is the most efficient of the methods since, for comparable validity, the rating method with four categories requires about one third the number of trials as that used for the binary

(yes-no) procedure. Also, Pastore and Scheirer (1974) acknowledge that the rating method is the most typical method for an ex post testing of the assumptions of SDT. Furthermore, Chapman and Feather (1971) indicate that when rating scale measures are employed, SDT becomes a versatile technique for the quantification of subjective reports.

In the following section, I present a discussion of the DVPM based on SDT, which also enables one to measure knowing that one knows.

#### The Decision Variable Partition Model

Ferrell (1972) proposed a signal detection measure of knowing what one knows, or the extent to which one can distinguish what one knows from what one does not. Also, Ferrell and McGoeys (1980) have proposed a model of subjective probability calibration in which knowing that one knows has a specific interpretation. A general description of the model, called the Decision Variable Partition Model (DVPM), is presented below.

#### General Description of the Model

Assuming that answers to questions can be classified as right or wrong, the DVPM model describes how to measure a person's ability to distinguish right from wrong answers. It models the task of giving one's

subjective probability that one has rightly performed a task or appropriately responded to a stimulus as consisting of two parts: (a) attempting to detect whether or not the response is correct, and (b) assigning a numerical probability value on that basis.

The detection part of the task, the model assumes, can be described by a signal detection model. A value  $x$  of a scalar decision variable  $X$  is generated, in a manner that depends on the task structure, from a joint consideration of the stimulus and response. Ideally,  $X$  is monotone increasing with the posterior probability that the response was actually correct. Called "indicated subjective certainty" by the authors,  $X$  is a random variable having a different probability density when the response is correct [ $f(x/C)$ ] than when it is not correct [ $f(x/\bar{C})$ ]. The individual is assumed to use a cutoff value on  $X$  to decide whether one is correct or not.

The numerical probability is assumed to be determined by partitioning the range of  $X$  into  $(m)$  intervals with a set of cutoff values  $\{x_i\}$ . Then the set of  $m$  allowable probability responses  $\{r_i\}$  is mapped onto the intervals, with higher values of  $r$  going to higher values of  $X$ . One then gives the response  $r_i$  corresponding to the interval into which the observed value  $x$  falls. Figure 4-5 illustrates this process.



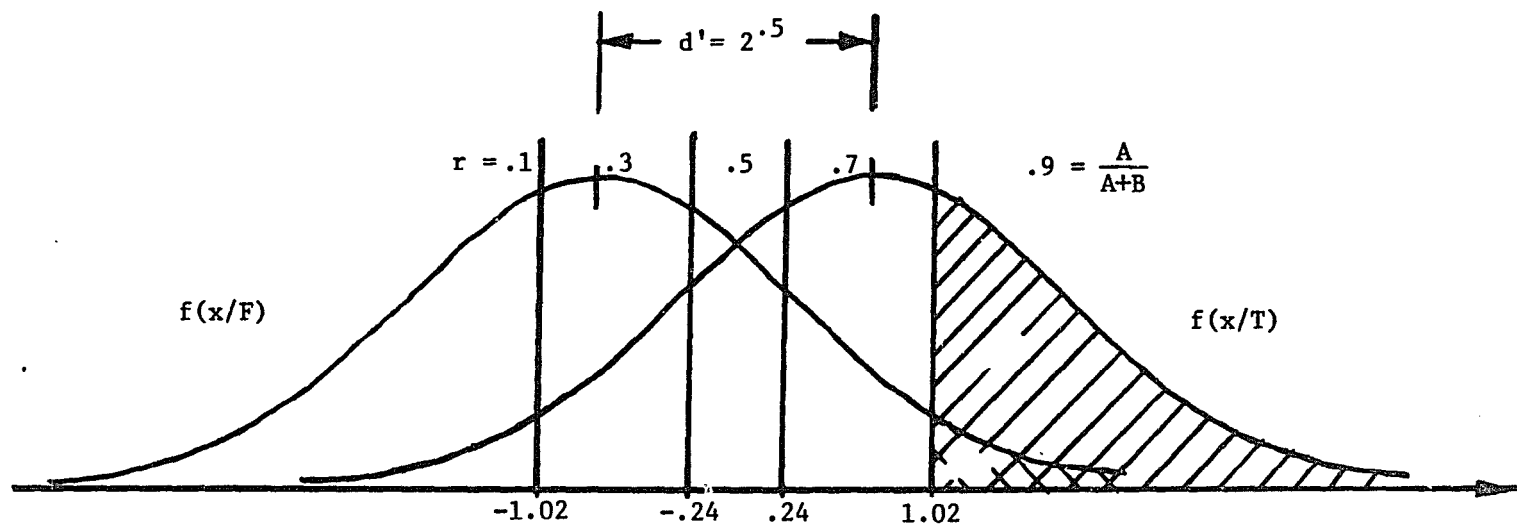


Fig. 4-5. The Partition on  $X = Y$  Which Gives Perfect Calibration for a YN(FR) Task When the Conditional Distributions of  $Y$  are Normal, Unit Variance with  $d' = 2.5$  and  $P(c) = .5$ .

### Independence of Knowing and Knowing That One Knows

The DVPM specifies the conditions under which knowing (knowledge) can be measured independently of knowing that one knows. For example, a subject is required to provide an answer to a question such as "What is absinthe" and then to give a probability  $r$  on the full range  $[0,1]$  that s/he is correct. This is the same as the signal detection task with a rating response. For this type of task, in which respondents supply their own answers, the detectability (say,  $A_c$ , which is the area under the curve) is a measure of how well the respondent can distinguish correct from incorrect responses, and this is potentially independent of the proportion of correct answers (say,  $P(c)$ ), which is a measure of knowledge). Hosseini-Ardehali (1981) reports the results of a study which supports the contention that  $A_c$  and  $P(c)$  are independent attributes ( $r = 0.0021$ ) under this stated condition.

It is not feasible to distinguish between  $A_c$  and  $P(c)$  in this study because the respondents did not supply their own answers. They were required to respond merely "Yes" or "No" to a question as to whether the stated book value is materially misstated, and then give a probability response  $r$  on the half-range  $[\cdot 5, 1]$  that their chosen alternative answer is correct. This task is identical to

the two-alternative forced-choice task in which a question is given with two answers (Yes or No), from which the respondent is to choose one. In this situation,  $A_c$  and  $P(c)$  are dependent (Hosseini-Ardehali, 1981).

#### Calibration and DVPM

In this study,  $A_c$  and  $P(c)$  are equivalent, hence the  $A_c$ s measured from the subjects' responses are interpreted as a measure of knowledge. To provide evidence relating to the degree of the subjects' sensitivity to their level of uncertainty, I provide evidence relating to the calibration of their responses.

In this model, experimentally determined calibration probabilities  $P(C/r_i)$  are the proportion of times one was correct when giving response  $r_i$ . The model's calibration is the proportion of responses that generate  $x$  values between  $x_{i-1}$  and  $x_i$  that are in fact correct, and can be determined from the probability of correct response  $p(c)$ , the cumulative distribution functions of  $X$  when the response is correct  $F(x/C)$  and not correct  $F(x/\bar{C})$ , and the partition  $\$x_i\$. The computation of calibration based on the model is described by the following equation:$

$$P(C/r_i) = \frac{[F(x_i/C) - F(x_{i-1}/C)] P(C)}{[F(x_i/C) - F(x_{i-1}/C)] P(C) + [F(x_i/\bar{C}) - F(x_{i-1}/\bar{C})] P(\bar{C})} \quad (4-10)$$

The partition is assumed to be determined by information obtained prior to the task (e.g., previous knowledge), which is not expected to change unless feedback about performance is given.

The DVPM appears to be a versatile tool for analyzing detection tasks in view of the following. First, DVPM provides a basis for predicting and evaluating the systematic effects of some relevant variables on calibration. In particular, it accounts for the effects of base rate and task difficulty on calibration. For example, the model predicts that there will be no effect of base rate on calibration in an experimental task in which subjects are to decide whether a proposition is true or false and, subsequently, to give subjective probability that the decision is correct.

The DVPM also predicts the effect of task difficulty (measured as proportion of correct responses [ $P(C)$ ]) on calibration. It states that for difficult tasks (i.e.,  $P(C)$  about .7 or less), subjects' judgments generally will be overconfident. Similarly, for relatively easy tasks, subjects' judgments generally will be underconfident. The model, therefore, provides a theoretical basis for predicting and evaluating the nature of calibration of the responses of this study's subjects.

Second, the signal detection aspect of DVPM enables one to determine (a) a nonparametric measure of

subjects' detectability as well as the type of decision errors which the subjects are more likely to commit, and (b) the decision rules employed. The latter provides a measure of the subjects' propensity to favor "sn" or "n", a detailed discussion of which I present in the next section.

#### The Criterion

In simple terms, the criterion refers to the decision rule employed by an observer in a discrimination task under conditions of uncertainty. It reflects the extent to which the observer favors one hypothesis over another. Criterion or bias suggests that, independent of the stimulus, not all responses are equally likely. As such, it is desirable to ensure that the criterion is not confounded with detectability. Craig (1979, p. 71) notes that a basic requirement for measuring performance in SDT is that detectability be independent of the criterion employed, since otherwise the former cannot be held solely to reflect the observer's ability to distinguish between the two events SN and N. One other reason for measuring the criterion in SDT tasks is to enable the researcher to determine whether one experimental treatment caused the observer(s) to favor, say, SN to a greater extent than in other experimental treatments.

The type of bias employed by an observer is determined by the subjective probability of item under consideration and upon the utility to the observer of the various possible decision outcomes. The concern with the decision rule is to provide answers to the following types of questions: Given an observation, what response alternative should be chosen? What is a good choice? How can one analyze these choices? These questions can be better answered by means of an example (see Green, 1960).

Assume a set of observations, each observation ( $X_i$ ) [e.g., a sealed package] represented by three numbers ( $X_i = [x_1, x_2, x_3]$ ) (e.g., length, width, and depth), and that there are two hypotheses  $H_1$  and  $H_2$  (e.g.,  $H_1$  means the package contains a toy car, and  $H_2$  means the package contains an animal) about the observations. Given an observation, a subject is to decide whether the observation is an instance of  $H_1$  or  $H_2$  (i.e., given the measurements of a sealed package, guess whether it contains a toy car or an animal).

If someone were to make a decision about a particular observation, as above, s/he would probably guess it was  $H_1$  if the probability of that observation was greater for  $H_1$  than for  $H_2$ . This statement is called the decision rule. It is usually defined in terms of a likelihood ratio (others call it the "odds"), which is the probability that a particular observation resulted from,

say, H1, divided by the probability that it resulted from H2. It can be represented by the following equation:

$$B(x_1, x_2, x_3) = \frac{P[H_1(x_1, x_2, x_3)]}{P[H_2(x_1, x_2, x_3)]} \quad (4-11)$$

The likelihood ratio is a number, not a probability, which in this example is a function of three variables ( $x_1, x_2, x_3$ ). Hence, we have taken an observation which is specified by three values ( $x_1, x_2, x_3$ ) and related it to a single variable  $B(x_1, x_2, x_3)$ . This transformation was performed because not only is the decision based on a single dimensionless number, but also one can always make "optimum" decisions when the likelihood ratio is used. For example, in equation (4-10), a respondent will choose H1 if  $B(X)$  is  $> 1$ . The number "1" is called the criterion or, more precisely, a likelihood-ratio criterion.

A decision procedure is "optimum" only if it best attains some specified objectives. Some such objectives in a judgmental task might be, but are not limited to: (1) maximization of the expected value of decisions, (2) minimization of risk, (3) estimation of a posteriori probability, or (4) maximization of the percentage of correct decisions. The literature indicates that the likelihood ratio criterion is optimum for all of the above objectives. Generally, an observer can make optimum decisions with the criterion defined for each of the

following situations (see Sheridan and Ferrell, 1980, McNicol, 1972):):

a) When  $P(SN) = P(N)$  and the observer is indifferent to the costs of decision errors and the rewards of correct decisions, then  $B^*(x) = 1$ . (4-12)

b) When  $P(SN) \neq P(N)$  but the observer is indifferent to the reward matrix, then  $B^*(x) = P(SN)/P(N)$  (4-13)

(c) When  $P(SN) = P(N)$ , but there exists different payoffs for correct versus incorrect decisions, then

$$B^*(x) = (VnN + CnS)/(VsS + CsN) \quad (4-14)$$

where  $VsS$  = value of making a hit,

$CsN$  = cost of making a miss,

$CnS$  = cost of making a false alarm, and

$VnN$  = value of making a correct rejection.

(d) If in (c)  $P(sn) \neq P(n)$ ,

$$B^*(x) = \{[(VnN + CnS)][P(n)]\} / \{[(VsS + CsN)][P(s)]\} \quad (4-15)$$

If the observer receives some information, say  $S1$ , prior to making a decision, the likelihood ratio based on prior probabilities, i.e.,  $P(H1)/P(H2)$ , is merely replaced by the posterior likelihood ratio

$$\frac{P(H1/S1)}{P(H2/S1)} = \frac{[P(H1)][P(S1/H1)]}{[P(H2)][P(S1/H2)]} \quad (4-16)$$



The defined criterion for each situation constitutes a benchmark for evaluating the appropriateness of the decision criterion used by an observer in a signal detection task.

The decision criterion defined in equation (4-15) is particularly useful, since it enables one to infer unambiguously the observer's preference between the different values of correct decisions and the costs of incorrect decisions.

An example will make this clear. Note that equation (4-15) can be written as follows:

$$B^*(x) = \frac{[V_nN + C_nS][P(N)]}{[V_sS + C_sN][P(SN)]} \quad (4-17)$$

which further can be defined as

$$B^*(x) = K \frac{[P(N)]}{[P(SN)]} \quad (4-18)$$

where  $K = (V_nN + C_nS)/(V_sS + C_sN)$

$K$ , the first term on the right side of equation (4-17) is a constant which depends on the reward matrix. It is the ratio of two quantities called regrets (Sheridan and Ferrell, 1981, p. 356). The numerator is the difference between what the observer would get for responding correctly and what he would get for responding incorrectly when  $N$  obtains (that is, the regret for incorrectly saying "sn"). Similarly, the denominator is

the regret for incorrectly responding "sn". It is apparent that  $K > 1$  if and only if the regret for incorrectly saying "sn" is greater than the regret for incorrectly saying "n", while  $K < 1$  if and only if the regret for incorrectly saying "sn" is less than the regret for incorrectly saying "n". When  $K = 1$ , it means the observer is indifferent to the consequences of his/her decision outcomes.

Assume, for example, that  $P(SN)$  is 0.4, and the experimentally determined measure of bias, say,  $B(x)$ , is 1.40. To determine the observer's preference between the decision outcomes, one calculates

$$B^*(x) = K[.6/.4] = 1.5K \quad (4-19)$$

from which  $K = 0.933$ . This  $K$  value suggests that the observer would have less regret associated with incorrectly responding "sn" than incorrectly responding "n". In the context of this study, this scenario suggests that the auditor will prefer to slate a materially correct account book value for intensive audit rather than incorrectly accept a materially misstated account balance.

#### Calculating the Criterion

In SDT, the method of calculating the criterion depends on the assumed underlying distributions for the events SN and N. In the equal variance Gaussian model, the criterion is independent of  $d$  and, according to

Pastore and Scheirer (1974, p. 947), is defined by the following equation:

$$B = [f(b/1)]/[f(b/2)] \quad (4-20)$$

where  $f(b/i)$  is the height of the probability density function for class  $i$  at the criterion or boundary  $b$  between the two response classes. Pastore and Schierer (1974) also note that  $P(sn/SN)$  and  $B$  are monotonically related.

The statistical program used for analyzing the subjects' responses in this study provides a measure of the subjects' overall criterion for classifying the stimuli either as SN or as N. The measure, the details of which are discussed in Chapter 7, also is based on the likelihood ratio. The measure of bias for each experimental condition is compared with the  $B^*(x)$  derived from the experimental data, as a basis for determining the subjects' preferences between the costs associated with their decision errors.

#### PAR As A Signal Detection Task

AR essentially involves a comparison of relationships amongst data whereby, through an evaluation of available evidence, the auditor decides whether the account book value is fairly presented, given his/her expectations of what the book value should be (or what the range of book values might reasonably be expected to be).

In essence, PAR enables the auditor to identify at an early stage of an audit the account items that are most likely to be misstated and, consequently, on which more audit efforts should be allocated. The accuracy of judgments in such tasks depends on the auditors' ability to discriminate those account items which are materially misstated from those which are fairly presented. Viewed in this manner, PAR is analogous to a detection task. Hence, a model based on signal detection theory, which is concerned with evaluating an observer's ability to discriminate between classes of events, is an appropriate tool for analyzing auditor judgments in this study.

Although signal detection theory (SDT) is an outgrowth of studies into problems of statistical analysis of radar signals, it has been successfully employed in analyzing discriminability in many other contexts, including medical diagnosis (Lusted, 1971); memory (Banks, 1970); criminal justice process (Pease, Tarling, and Meudell, 1977); the performance of a group of on-line inspectors in an industrial setting (Drury and Addison, 1973); and cost-variance investigation decisions in accounting (Brown, 1981). Also, Blocher and Moffie (1982) recently have suggested the use of the signal detection model for analyzing accounting and auditing judgments.

The concepts of SDT discussed earlier can be described in the context of AR as follows. The external

auditor is the detector (or observer); the stimulus is the reported book value, such that SN means the book value is materially misstated, while N means that the book value is fairly presented. Accordingly, 'sn' means that the auditor believes that the book value is materially misstated, while 'n' means s/he believes that the account book value is fairly presented.

In terms of Statement of Auditing Standards No. 39 (Audit Sampling), (SN,n) is equivalent to the decision error of incorrect acceptance of a materially misstated book value, while (N,sn) is equivalent to the decision error of incorrect rejection. In the AR context, the hit rate refers to the probability of rejecting a materially misstated book value, while the false alarm rate refers to the probability of rejecting a fairly presented book value. Detectability in the signal detection sense is used as an approximate measure of auditor's knowledge. The area under the curve, which is the assumption-free measure of detectability described earlier, is used to compute auditors' detectability.

The decision criterion defined in (4-15) is of direct relevance to this study, since one expects an auditor to attach different costs to the incorrect decisions in the decision matrix. For example, the literature (e.g., Robertson, 1979, p. 343) suggests that the error of incorrect rejection of a fairly presented

book value is not considered to be as serious as an error of incorrect acceptance of a materially misstated book value. By implication, the auditor perceives the cost of incorrect acceptance to be greater than the cost of incorrect rejection. Equation (4-15), therefore, is used to provide an unambiguous measure of auditors' preferences between the costs of their decision errors. An analogous comparison of the concepts of SDT in the PAR context is illustrated in Figure 4-6.

In the next chapter, I describe the methodology of the study's experiment, and indicate the SDT-based performance measures used to provide evidence bearing on the research issues addressed by this study.

<u>STD</u>	<u>PAR</u>
1) Detector	External Auditor
2) Stimulus	The dollar amount of account book value evaluated, and classified as N or SN, by the auditor, in light of his/her substantive knowledge and other information provided in the experimental materials.
3) Signal-plus-Noise (SN)	An account item which actually is materially misstated.
4) Noise only (N)	An account item which actually is fairly presented.
5) Detectability	A measure of Auditor's ability to correctly recognize account book value which is materially misstated.
5) Judgment Bias	Implicit Loss Functions (i.e., motivational factors influencing auditor's decision to classify an account item as SN or N).
7) Response Criterion	Auditor's subjective standard (or critical value) for deciding whether an account book value is N or SN.
8) Prior SN probability (or Base Rate)	The proportion of Account Items which actually are materially misstated.
9) Number of Trials (or Observations)	Number of account items on which each auditor provided a response.

Fig. 4-6. Signal Detection Theory (SDT) and Preliminary Analytical Review (PAR): a Comparison of Concepts

## CHAPTER 5

### METHODOLOGY

To accomplish the objectives set forth in Chapter 3, I performed a research study in which practicing auditors provided responses to a set of questions relating to selected account balances for each of two experimental cases: (1) one case based on one client firm for which the internal control system was adjudged strong, and (2) another case based on another client firm whose internal control system was adjudged relatively weak. The subjects' responses were analyzed (1) in the aggregate, (2) by state of internal control, and (3) by functional level. An overview of the experimental design for this study is presented in Figure 5-1.

#### Scope of the Study

The study focuses on the evaluation of auditor judgments at the initial planning stages of an audit (i.e., preliminary analytic review), for reasons stated earlier. Furthermore, others (e.g., Holder and Collmer, 1980, p. 31) have noted that substantive AR usually is



		INTERNAL CONTROL		Number of Subjects
		Weak	Strong	
FUNCTIONAL LEVEL	Senior			20
	Manager			8
TOTAL				28

Fig. 5-1. Experimental Design

applied when preliminary ARs are not applied in the planning phase and/or material adjustments are made to financial statements during audit in light of factors or evidence not anticipated at the onset of the audit. They also suggest that, to enhance audit effectiveness and efficiency, the nature, timing and extent of substantive tests be based on both the evaluation of internal accounting control and the application of (preliminary) AR procedures during the planning process (stage) of an audit.

In practice, audit adjustments are made only if the net effect of errors identified in account balances during an audit have a material effect on, for example, net income. Therefore, in the case of compensating errors, significant errors identified in individual accounts may not result in an audit adjustment. For example, assume that the net income of a client is \$10 million. If, for example, cost of sales is overstated by \$3 million, and selling expenses are understated by \$2.999 million, the net effect on net income is only \$1,000, which the auditor may consider immaterial and, hence, no audit adjustment may be proposed.

The auditor, however, normally would devote audit efforts toward identifying the nature and magnitude of the errors in each account item before he could estimate the net effect of all the errors on net income. To enhance

audit effectiveness and efficiency, therefore, the auditor's objective should be the identification of each account book value which is materially misstated. Hence, the relevant task for the auditor, for the purpose of this study, is to identify account balances which, evaluated independently of any compensating errors in other accounts, are misstated.

### The Subjects

The subjects for the study were twenty-eight practicing auditors employed by two of the "Big Eight" public accounting firms with offices in Phoenix, Arizona and Los Angeles, California. Having agreed to provide subjects for the study, each public accounting firm selected the specific participants. The subject selection is, therefore, nonrandom; willingness to participate and availability were the selection criteria.

To ensure a reasonable degree of substantive experience, I requested that subjects have about two years of experience in analytical review judgments. As a result, only senior- and manager- level auditors, who normally are directly involved in analytical review judgments, participated in the study. The participants' length of service ranged from 18 months to 97 months, with an average of 47 months. Also, they had worked on an average of 32 audit engagements, with a range of 4 to 100.

There are three measures of task-related expertise which I consider relevant for analyzing this study's data. These are (a) length of service in public accounting, (b) number of audit engagements, and (c) functional level (i.e., manager or senior). I used functional level for analyzing the study's data in view of the following. First, the number of audit engagements is not a true reflection of experience because of the differences in the size of the clients each auditor might be assigned to. The auditors assigned to large clients typically record a lower number of audit engagements than those assigned to small clients. In fact, for this study's subjects, the degree of association between functional level and number of audit engagements, measured by eta (see Nie, et al., 1975) is only 0.67. Second, although length of service highly correlates with functional level ( $\eta = 0.88$ ), discussions with auditors in several CPA firms indicated that functional level is the sole criterion for participation in AR judgments. For example, the discussions revealed that staff auditors typically do not participate in AR judgments. Furthermore, AR-related judgment is a multi-stage process in which the senior's initial judgments are reviewed by the manager before a final decision is made. Hence, there seems to be an implicit assumption in practice that functional level is the best measure of expertise.

To motivate the auditor-subjects participating in the study, I solicited and obtained the subjects' employers' sanction of the research study. Furthermore, the study was performed on the employers' premises during regular working hours. This served not only to remind the subjects that the study had been officially approved by their respective employers, but also to enhance the perceived importance of the study.

#### The Case Studies

To enhance the realism of the study, two experimental cases were developed from data provided by a national public accounting firm on two independent audit clients in the electronics industry. Deciding what information to provide was not an easy task, since auditing literature does not provide an adequate guide regarding the information which should be gathered and evaluated for PAR judgments. The following considerations, therefore, guided the choice of the information provided in the case studies.

As indicated earlier, the auditing literature notes that AR involves a comparison of relationships among data. On this basis, I consider it appropriate to provide audited financial statements for at least two previous years to provide evidence on the financial statements' trends for the two firms used in the case studies.

However, other financial data provided (e.g., financial ratios) were determined by data availability and the need to provide comparable data sets for the two cases.

Auditing theory suggests that good performance in PAR tasks requires a familiarity with, and an understanding of, the nature of the client firm's industry, business operations, accounting procedures, and other qualitative factors such as the quality of accounting personnel. GAAS also indicate that the extent of substantive tests required to constitute sufficient evidential matter under the third standard (of field work) may properly vary inversely with the auditor's reliance on AIC.

These observations suggest the need to provide information relating to the nature of operations and industry of the firms whose data were used in the case studies. They also suggest the need to provide information relating to the quality of each firm's AIC system.

To further enhance the reasonableness of the information provided in the case studies, an audit manager in the public accounting firm which provided the data assisted in determining what constitutes a "reasonable" information set typically available at the onset of an audit, particularly in respect of the two client firms whose data were used for the study. Consequently, each

case presented the following information: (1) brief background information on the electronics industry; (2) general, operating and financial data on each client firm for 1979 and 1980; and (3) a description of each firm's internal control system. Appendix A shows the experimental materials containing these data.

Criteria for Selecting and  
Classifying Account Items

SDT largely has been applied to sensory and auditory detection tasks in which the SN and the N items are clearly distinguishable. However, in other applications like auditing, there is no unequivocal basis for classifying the stimuli into SN (i.e., materially misstated) or N (i.e., fairly presented) categories. In such applications, the classification has to be made on the basis of arbitrarily determined criteria. For example, in a study in which SDT was applied to a compliance testing experimental task, Blocher and Moffie (1982) assumed that a population with an arbitrarily (emphasis added) low error rate is acceptable, while the high error rate population is unacceptable.

While this arbitrariness may be an indispensable feature of the application of SDT to accounting and auditing judgment tasks, it is noteworthy that an arbitrary cutoff value (such as error rate in compliance testing) may have a differential impact on the performance

of each subject, given that what constitutes a low or high cutoff value may vary significantly between individual subjects. To alleviate the potential negative impact of this arbitrariness, it might be desirable to (1) identify, in whatever way possible, a reasonable cutoff value which is acceptable to most of the subjects or (2) perform a sensitivity analysis of the variations in the cutoff rate on the subjects' performances.

In the AR context, the presence or the absence of an audit adjustment could be one criterion for classifying an account item into SN (i.e., as materially misstated) or as N (i.e., as not materially misstated). This criterion is, however, not unequivocal, in view of the following. First, the audit process itself determines whether the "need" for an adjustment would be detected. Therefore, materially misstated book values still may go undetected. Second, the decision to record or not to record an audit adjustment in light of a detected error is a matter of individual auditor judgment. Therefore, two or more auditors may reach different conclusions on the same account item. Furthermore, in practice, auditors decide to slate or not to slate an account item for intensive audit for reasons other than the suspected presence of material errors alone, such as the perceived significance of an account item to the financial statement as a whole.



The discussion above indicates that the basis for classifying account items as N or SN necessarily will be arbitrary. Hence, it will be appropriate to provide evidence regarding the sensitivity of the subjects' responses to plausible alternative classification approaches that may be adopted. Before discussing the three classification alternatives adopted in this study, I discuss the nature of the errors reported in the accounts of each client firm used for this study.

For the ABC company, in which the AIC system is adjudged strong, the auditor detected errors in four account items. The errors in two of these accounts would have overstated net income by about \$80,000, while the errors in the other two accounts would have understated net income by about \$40,000. Since the net effect of these compensating errors would have been an understatement of net income by about 4%, a proportion which the auditor considered immaterial, no actual audit adjustments were booked for ABC company.

The auditor also detected errors in five accounts of company XYZ, in which the AIC is considered relatively weak. Actual audit adjustments were booked in only two of these accounts. The others were not booked as audit adjustments, since the dollar amounts involved were considered by the auditor to be immaterial.

In view of the above situation, I consider the following three approaches appropriate for classifying the study's account items as N or SN. First, I categorize as SN all the account items in which misstatements (both significant and insignificant) were detected in each company. This approach has a basis in practice. For example, some CPA firms control only for undetected errors and, therefore, they record any error detected during an audit. That is, all detected errors are considered significant and are hence treated as if they were all audit adjustments.

Second, I consider as SN all the XYZ accounts for which actual audit adjustments were booked. Correspondingly, for company ABC, I categorize as SN the account items with the larger dollar amount of error (i.e., \$80,000). The decision is based on the assumption that the dollar magnitude of the error, considered independently of any compensating errors in other accounts, is large enough to warrant detailed investigation by the auditor.

Third, as in the second approach, I again categorize as SN all the XYZ accounts for which audit adjustments were booked. However, for company ABC, I classify as SN the accounts with the lower dollar amount of error (i.e., \$40,000). This categorization is based on the assumption that, regardless of compensating errors

detected in other accounts, the auditor may not overlook any error which may overstate net income.

The eventual categorization of the accounts under the three approaches described above is shown in Figure 5-2. Note that the effect of the classification alternatives is to reduce the prior signal probability  $[P(SN)]$  from 0.4 and 0.5 for ABC and XYZ respectively under the first classification approach to 0.2 for both ABC and XYZ under the other two classification approaches. The data analysis will provide evidence regarding the effects of these classification approaches on the subjects' performance.

#### Experimental Task

This study's experimental task belongs to the class of one-alternative, probability-correct tasks described by Smith and Ferrell (1981). The subjects were requested to identify the current year's (1981) account items which they believe to be materially misstated.

As indicated earlier, I intended to provide evidence regarding the nature of calibration of the subjects' judgments. But, as Lichtenstein, et al (1982) indicate, it is impossible to determine whether the judgment of an individual (or a group of individuals) is well calibrated when confidence is expressed on a rating scale. The subjects were, therefore, requested to state

ACCOUNT ITEMS	I		II		III	
	A B C	X Y Z	A B C	X Y Z	A B C	X Y Z
Sales						
Cost of Sales	X	X	X			
Income Tax Provision						
Inventory	X	X	X	X		X
Accounts Receivable						
Allowance for Doubtful Accounts	X				X	
Bad Debt Expense	X				X	
Accounts Payable		X				
Plant, Property & Equipment		X				
Depreciation Expenses		X		X		X

Note: I = Account Classification Approach Number One  
 II = Account Classification Approach Number Two  
 III = Account Classification Approach Number Three

ABC = Strong Internal Control System Situation  
 XYZ = Weak Internal Control System Situation

x refers to those account items classified as SN (i.e., materially misstated), based on the errors documented in the audit working papers.

Fig. 5-2. Categorization of Account Item as N or SN  
 By Account Classification Approach.

subjective probabilities on a half-range [.5, 1] probability scale regarding their degree of belief in the correctness of their responses. This half-range probability scale has been suggested by Smith and Ferrell (1981) as appropriate for this class of experimental tasks.

Each subject provided responses to ten account items for each case. The presentation of the cases and the individual account items within each case was randomized to control for order effects.

Conclusions Versus Decisions in  
Experimental Tasks: A Distinction

The experimental task (described above) requires that the subjects state whether a stated book value is (is not) materially misstated, and also to provide a subjective probability regarding the correctness of the response. However, there is a potential difference between a judgment criterion and an action criterion. In other words, the criterion for deciding whether an account item is materially misstated may differ from the auditors' criterion for slating that account for intensive audit.

This distinction between the judgment criterion and the action criterion is analogous to the distinction between reaching a conclusion and making a decision noted by Tukey (1960). He stated that a conclusion is a statement which is to be accepted as applicable to the

conditions of an experiment or observation unless and until unusually strong evidence to the contrary arises. These conclusions, he said, are established with careful regard to evidence, but without regard to consequences of specific actions in specific circumstances.

Decisions, on the other hand, are based not only on an evaluation of the available evidence, but also on the consequences of alternative courses of actions open to the decision maker. That is, a decision entails the weighing of both the evidence concerning the relative merits of two (or more) alternative hypotheses and also the possible consequences of various actions, from which one decides that a particular course of action is the most appropriate to take under the given circumstance. Hence, a decision involves an attempt to choose the best risk in a given uncertain situation.

The judgment criterion in this study is similar to the process of reaching a conclusion, while the action criterion corresponds to the process of making a decision. These are two separate and important aspects of judgments which earlier studies have neglected. The relevance of this distinction in the context of this study is elucidated by the following.

Discussions with practicing auditors indicate that even though the auditor might believe an account book value to be fairly presented, s/he may decide to slate it

for intensive audit efforts. Similarly, s/he may decide not to slate for intensive audit an account item perceived to be materially misstated. These decisions are, usually, influenced by considerations such as the perceived importance of the account item, or whether the incremental benefit of providing intensive audit efforts to ascertain the correctness of a stated book value justifies the additional audit cost involved.

It is not possible to state a priori what the effect of each classification strategy on the subjects' responses would be. If the auditors' action (i.e., decision) criterion is more (less) strict than their judgment (i.e., conclusion) criterion, then the auditors are more (less) likely to commit the error of erroneously rejecting (accepting) fairly presented (materially misstated) book values. In either case, the auditors' decision rules for the action (decision) criterion may differ significantly from their decision rules for the judgment criterion.

To provide evidence relating to this idea, the auditors also were requested to indicate the account items on which they would plan to provide intensive audit efforts in light of their responses and other subjective considerations, such as, for example, the perceived importance of the account item in relation to the financial statements taken as a whole. However, given the

absence of a suitable statistical method for analyzing the point estimates (index) of auditors' decision rules by task criterion, only a qualitative description and evaluation of the significance of these indices could be made, as will be shown in Chapter 7.

#### The Pilot Study

Two senior- and manager level auditors employed by one of the "Big Eight" public accounting firms participated in the pilot study. They were similar in terms of functional level to the participants in the actual study. They also performed under experimental conditions similar to those used in the actual study.

The purpose of the pilot study was (1) to determine the clarity of the experimental materials and the adequacy of the limited information provided, and (2) to identify and rectify any potential problems the subjects might have with the experimental materials.

#### Administration of the Study

Subsequent to the pilot study, I administered the experiment at the subjects-auditors' offices. Initially, the participants were oriented through a brief discussion of the objectives and the focus of the study. Limited training also was provided by "walking through" an example of the experimental task and experimental materials.



Having answered the questions posed by the participants, I provided the actual experimental materials (shown in Appendix A) to them. I was present during the performance of the experimental task to answer questions and to assist the subjects as necessary.

Kadane and Lichtenstein (1982) indicate that to evaluate the calibration of judgments for nonexchangeable items (such as account items used in this study), the subjects should be provided with feedback after each response. Providing feedback to the subjects in this manner would have resulted in a significant departure from audit practice, since feedback is normally not available to auditors at this stage of an audit. Hence, no I did not provide feedback to the subjects of this study.

The subjects were given two and one-half hours to perform the experimental task, and this time limit appeared to be adequate. However, I did observe that, in general, the seniors took more time than the managers to complete the experimental task.

No interaction between the subjects was permitted. Furthermore, the subjects were separated physically and instructed not to pay attention to the others in the room.

Having completed the experimental task, the subjects were requested to provide responses to a set of background questions relating to the following: (1) present position; (2) professional qualification(s); (3)

college educational level; (4) number of years and variety of work experience; (5) number of college courses taken in probability and statistics, including the number of hours of in-house statistical training received, and (6) specification of the materiality threshold, if any, used by the subjects for their judgments.

Discussions were held with each participant after the exercise to evaluate their interest in the experiment. Without exception, they expressed satisfaction with having participated in the study. They also found the experiment interesting and worthwhile, and each one of them requested a copy of the study's findings.

#### Research Issues Addressed

The study provided data useful for evaluating several aspects of the nature and characteristics of auditor judgments in PAR tasks. In this section, I present a detailed discussion of the specific research issues addressed in this study.

##### Research Issue Number One:

What is the detectability  
of Auditors' Judgments?

Detectability is a measure of the extent to which an observer can appropriately distinguish the two events SN and N independent of any judgment criteria (or biases)

employed. In the context of this study, it provides a measure of the accuracy of the subjects' judgments. The higher the index of detectability, the higher the accuracy of their judgments. As which, as indicated earlier, is a nonparametric measure of detectability, is used in this study.

A computer program written by Grey and Morgan (1972) was used to analyze this study's data. The program, which assumes that the events are normally distributed, uses the maximum-likelihood estimation method to derive the parameters of interest given data from detection tasks.

This program is appropriate for analyzing the data for this study because, as Ogilvie and Creelman (1968) have shown, the conventional least-squares curve fitting procedures are inappropriate for ROC plots, since both axes represent dependent variables and are both subject to error. The normal distribution model also is considered appropriate, since Bush (1963, p. 454) has shown that the normal distribution model compares favorably to other eligible models (e.g., the logistic model) for analyzing this type of data. Furthermore, Luce (1963, p. 61) has indicated that ROC curves generated by the logistic and normal models are virtually indistinguishable.

To enhance the reader's understanding of how the evidence bearing on this study's research issues was

generated, I present in Appendix B (a) an overview of the statistical model suggested by Grey and Morgan (1972) used in this study, and (b) the procedures used to analyze the subjects' responses.

Ideally, Ac could be regarded as a measure of auditor's detectability in this study. However, unlike sensory and auditory tasks to which SDT largely has been applied, in this experiment there is neither a well-defined information set nor are the stimuli controllable by the experimenter. Furthermore, auditors normally have more information available in practice than was provided in the experiment. It appears, therefore, inappropriate to regard Ac strictly as a measure of auditors' knowledge in PAR tasks since the Acs reported for the subjects in this study would likely be underestimated to an extent which is, however, unknown.

#### Research Issue Number Two:

What Type of Response Biases Are Exhibited by the Auditors?

The answer to this question provides an overall quantitative index of the judgment criteria (or biases) employed by the subjects in the aggregate and both by the (a) state of AIC and (b) functional level. It provides an indication of the subjects' propensity to respond "sn" more than "n", or vice versa.

The computer program used for this study also computes the measure of response bias for each cutoff point (say,  $Z_m$ ) on the probability scale. As explained in Chapter 4, the measure of response bias (say, Beta) usually is defined in terms of the likelihood ratio, that is, the distribution of SN at each cutoff point divided by the distribution of N at the same cutoff point. For example, the beta can be defined as follows:

$$\text{Beta} = f_{sn}(Z_m)/f_n(Z_m). \quad (5-1)$$

Of interest, however, is the likelihood criterion around the 0.5 response category, since it provides a measure of the subjects' criterion for classifying the stimuli either as SN or as N. As indicated earlier, each data set was collapsed into ten categories. Therefore, to derive an overall measure of bias for each data set, I used the average of the  $Z_m$  values for the fourth and fifth cutoff points to calculate the Beta criterion. That is,

$$\text{Beta}' = f_{sn}(Z_m')/f_n(Z_m'), \quad (5-2)$$

where Beta' refers to the overall measure of response bias for each data set and  $Z_m'$  refers to the approximate value of the median cutoff point. Beta > 1 suggests that the subjects were more inclined to respond "sn", while Beta < 1 suggests an inclination towards responding "n". A Beta of 1 indicates an equal disposition (indifference) by the subjects towards classifying the stimuli either as SN or as N.

The optimum Beta value (i.e.,  $B^*(x)$  for equation 4-12 in Chapter 4) also is derived for each data set from which  $K$  (defined earlier) was determined. The value of  $K$  provides a basis for determining the subjects' preferences between the consequences of their decision outcomes, as demonstrated in Chapter 4. There, it was shown that  $K > 1$  ( $K < 1$ ) indicates that the observer believes the regret for incorrectly responding "sn" is greater (less) than the regret for incorrectly responding "n".  $K = 1$  means the observer is indifferent to the consequences of the decision outcomes.

A virtue of this measure is that it enables one to avoid the problem of identifying or measuring the specific utility function employed by the observer (or subject). Thus, almost all the assumptions about the values in the payoff matrix are neutralized or cancelled by this measure (Licklider, 1964, p. 113).

As suggested earlier, the measure of judgment criteria employed provides a "validity" check on the type of decision errors which auditors are likely to commit. If the subjects have a propensity to respond "sn", there is a likelihood of an increase in the false alarm rate, suggesting that the auditors are more likely to commit errors of incorrect rejection of stated account balances. A propensity to respond "n" results in a likelihood of increasing the number of misses, in which case auditors

will be more prone to an incorrect acceptance of materially misstated account balances.

Research Issue Number Three:

What is the effect of Task  
Criterion on Auditors' Performance?

As indicated earlier, Tukey (1960) has made a distinction between conclusions and decisions in experimental tasks. The distinction suggests that the auditor should be more conscious of the consequences of his/her judgments while making decisions than when reaching conclusions.

Evidence bearing on the effect of this distinction should be of interest. For example, if significant differences were found in the subjects' decision rules by task criterion, this will indicate that their beliefs may not be independent of their preferences. Such a finding may have implications for the application of Bayesian principles to the audit decision process, since the principles require that the judges' beliefs be independent of their preferences. Evidence bearing on this research issue will indicate the impact of this distinction of the subjects' decision rules. Unless the subjects are risk neutral, one expects, apriori, that their decision rules for the action criterion will be more strict than their decision rules for the judgment criterion.

Statement of Auditing Standard (SAS) No. 39 suggests that there are two types of decision errors concerning stated book values which auditors might commit: (1) the error of incorrect acceptance, and (2) the error of incorrect rejection. stated book value. SAS No. 39 also states that the error of incorrect acceptance relates to the effectiveness, while the error of incorrect rejection relates to the efficiency, of the audit. Hence, an evaluation of the effect of task criterion on decision errors will, indirectly, provide evidence relating to the effect of task criterion on the effectiveness and efficiency of the audit.

I used the parametric test for differences in proportions between matched samples, suggested by several researchers [for example, McNemar (1949); Cochran (1950); Glass and Stanley (1970)], to provide evidence relating to this research issue. It is appropriate for evaluating the significance of the differences between the values of the cells in contingency tables, like the one shown in Table 7-4. The details of this test are discussed in Chapter 7.

#### Research Issue Number Four:

What is the effect of the state of AIC on Auditors' performance?

When internal controls are adjudged strong, the auditor might expect more account book values to be fairly



presented than if the internal control is adjudged weak. This idea is derived from auditing literature which, as discussed earlier, assumes that the existence of a satisfactory AIC reduces the probability that material errors in the accounts will occur and go undetected. In that case, the extent of substantive tests may be reduced.

A potential problem, however, relates to the reasonableness of an auditor's reliance on a given internal control situation. For example, the auditor may overrely on an internal control adjudged strong, or underrely on an internal control adjudged weak. SAS No. 39 indicates that overreliance on internal control can lead to an incorrect acceptance of an account balance, while an underreliance may lead to an incorrect rejection of an account balance. One expects, therefore, that the auditor would slate more account items for intensive audit when the internal control is adjudged relatively weak. An answer to this research question was provided by calculating (1) Ac, and (2) the decision errors auditors are more likely to commit by AIC environment.

#### Research Issue Number Five:

What is the effect of functional level on auditors' performance?

I have discussed in Chapter 3 the basis for an expectation that audit managers should outperform audit

seniors in PAR tasks. In addition, Taylor and Glezen (1979) suggest that the accuracy of auditor judgments in preliminary AR tasks requires a reasonable understanding of the client firm's general business and industry conditions, its peculiarities, and its accounting policies and procedures. Furthermore, in practice, AR involves a multi-stage process in which, for example, the judgment of the audit staff is reviewed by the supervising senior auditor, whose judgment is in turn evaluated by the audit manager in charge of the engagement.

By implication, the more experienced auditors are expected to possess greater expertise and should, therefore, be able to make more accurate judgments. The differences in Ac by functional level will provide an answer relating to this research question. Also, evidence regarding the effect of functional level on (a) type of decision errors likely to be committed, and (b) the preferred judgment biases will be provided.

#### Research Issue Number Six:

How Effective Are Auditors at Communicating Their Knowledge in Preliminary Analytical Review Tasks?

Subjective probabilities are used to indicate uncertainty about events of interest. A desirable property of such probabilities is that they be consistent with relative frequencies in the sense that the proportion

of actual occurrence of events which were assigned a given probability number should, in the long run, approach that number. Otherwise, the subjects' judgments will be miscalibrated, in which case they may be overconfident or underconfident. Such calibration's effects on audit effectiveness and efficiency have been noted earlier. For example, overconfident judgments may result in auditors' collecting less than adequate sample information on which to base their audit judgments. Similarly, underconfidence implies that auditors might collect more sample information than is required to make audit judgments.

In general, assessors either overestimate or underestimate their perceived degree of uncertainty in a given situation, either of which could affect the ability to effectively communicate their knowledge about the event(s) under consideration. As Ferrell and McGoey (1980) indicate, there are two aspects to knowing about one's ability to answer questions under uncertain conditions: (1) the capacity to distinguish correct from incorrect responses, and (2) the capacity to encode the distinction in a useful numerical form. The first of these is closest to the idea of "knowing how much they (the assessors) know"; the second is the effective communication of that knowledge and requires good calibration.

An appropriate measure of the capacity to distinguish is the detectability of correctness, in the signal detection theory sense. The answer to research question number one provided evidence bearing on this with respect to this study's subjects. To evaluate the subjects' capacity to effectively communicate their knowledge, I also evaluated the calibration of the subjects' judgments using the DVPM described earlier. If the subjects are able to communicate their knowledge effectively, they should be well calibrated. Otherwise, calibration may indicate overconfidence or underconfidence. In case of overconfidence, the subjects would have overestimated their capacity to detect account items which actually were materially misstated, with the opposite true in case of underconfidence.

The accounting literature suggests that the nature of auditor's subjective probabilities may be sensitive to the relative strength of internal control. For example, Solomon, Krogstad, Romney, and Tomassini (1982) report that auditors' prior probability distributions (PPDs) for account balances assessed for the stronger AIC system cases were less dispersed than the PPDs assessed for the weak AIC system cases. They also found that auditors' judgments were in closer accord when they faced the stronger AIC environment than when they faced the weaker AIC environment. This finding, Solomon, et al (1982)

indicate, provides preliminary evidence of auditors' sensitivity to the validity of their subjective judgments. Accordingly, they suggest a more direct investigation of the calibration of auditors' account balance PPDs. The DVPM also assumes that the partitioning of a subject's decision variable will depend to a large extent on his/her accumulated relevant knowledge prior to the task (e.g., previous experience). One expects, therefore, that these two factors should affect the calibration of auditors' probabilistic judgments in PAR tasks. Hence, the effects of the (1) relative strength of internal control and, (2) functional level on calibration were evaluated.

Finally, as indicated earlier, DVPM enables one to predict the effects of base rate and task difficulty on the calibration of the subjects' responses. If the proportion of correct responses is high (e.g., greater than .7), then one would expect the subjects' responses to show a tendency towards overconfidence. Similarly, one expects the base rate (i.e.,  $P(SN)$ ) to have no significant effect on the calibration of the subjects' responses.

#### Research Issue Number Seven:

What Types of Information do  
Auditors Require for PAR Judgments?

The evidence-collecting process for an external audit is generally thought to consist of three major

classes of evidence: (1) internal control evaluation and compliance tests, (2) tests of details, and (3) analytical review procedures (ARPs). As Mock, et al (1982) indicate, ARPs are probably the least well specified conceptually in the authoritative literature and within the audit judgment process. Specifically, ARPs are thought by practicing auditors to include a wide variety of auditing tasks, including the gaining of a general understanding of the client and its environment, the judgmental scanning of financial data, and the use of rigorous statistical models and tests. Also, Blocher, et al (1981) acknowledge that AP is not clearly defined, and that SAS merely sets forth a concept of AR which can be interpreted in many ways.

A review of the literature indicates that differences in the type of information used may affect the accuracy of, or be a potential source of variability in, auditor judgments. For example, Abdel-Khalik and El-Shesai (1980) indicate that the subjects' choice of information rather than their processing of chosen cues was the limiting factor in predicting the default on debt. Also, in recognition of the potential effect of differences in information search and choice behavior on consensus of auditor judgments, Mock, et al (1982) have suggested studies of information which auditors use for ARPs.

Prior research, however, has concentrated exclusively on aspects of information usage by auditors in experimental tasks. In such situations, the experimenter provides the types of information which s/he considers relevant to the given task. The intensity of usage of each type of information provided, and their effects on the accuracy of the participants' judgments, are then analyzed. For example, in a recent AR-related experimental study, Blocher, et al (1981) provided forty-four participants with trend analysis and operating data. They reported that thirty-three (75%) of the forty-four participants chose trend analysis, while only eight (19%) used the operating data in any significant way. The authors indicate that a simple reasonableness test using available operating data would have identified a material difference between reported payroll expense and the amount of payroll expense implied by operating data. But, they note, none of the participants detected this difference.

This approach, however, presumes that the experimenter knows either (a) the types of information that are really appropriate for the given task, or (b) the types of information which the participants would have required in practice to facilitate their judgments. These conditions are not likely to be satisfied,, given the unstructured nature of AR noted above. What is required,

therefore, is evidence regarding the types of information which auditors believe are relevant for their AR judgments, at least in a given situation.

To provide evidence bearing on this issue, I requested the auditor-subjects to specify, in a decreasing order of importance, the information items which they would have required in practice to facilitate their AR judgments with respect to each account item, irrespective of the information provided in the experimental materials. The relative importance of the information items specified for each account item was measured by the weighted average of the ranks assigned. The information with the highest average mean rank is considered the most important for each given account item. Also, to evaluate the degree of consensus among the auditors regarding the relative importance of each type of information, I calculated the coefficient of variation of the ranking derived for each type of information. When there is a high degree of consensus among auditors regarding the importance of an information item, the coefficient of variation should be very low. Perfect agreement would be indicated by a zero coefficient of variation.

#### Applicability of SDT to Groups of Subjects

SDT is ideally applicable to situations in which the experimenter controls the signals given to individual



subjects over short experimental sessions. Furthermore, the ideal SDT experiment requires a very large number of trials per subject. In this study, SDT was employed to analyze the performance of a group of subjects, each of whom was presented only a small number of trials because of (a) data limitations, and (b) the enormous cognitive demands which the evaluation and processing of information provided in the experiment would have placed on the subjects.

The applicability of SDT in this manner and the summation of data across groups of subjects has been attested to in the relevant literature. Angus and Daniel (1974) applied SDT to a marketing experiment in which a panel of judges was asked to rate the richness of 27 different ice cream products on a ten-point certainty scale. They found that SDT is an appropriate method for separating the observers' judgment criteria from their ability to perceive differences in richness, despite the small number of trials per subject.

Drury and Addison (1973) analyzed the records of the performance of a group of on-line inspectors of glass items over a ten-month period, using the SDT. All of the data refer to the total weekly performance of the inspectors in the group on all shifts and over a wide variety of faults. Therefore, "the data are radically different from the usual SDT data where the experimenter

controls the signals given to individual subjects over short experimental sessions" (p. 161). They found that the inspectors as a group behaved as SDT predicts, and hence conclude that "SDT, derived from carefully controlled experiments on individual subjects over very short periods of time, provides a useful description of the performance of groups of industrial inspectors over considerable periods of time" (p. 167).

Chapman and Feather (1971) studied the effects of deep muscle relaxation in a systematic desensitization context. Two groups of (student) subjects were examined: one set of 15 subjects imagined scenes while under deep muscle relaxation, while another set of 15 subjects were not relaxed while imagining scenes. Each subject rated every phobic scene imagined with regard to the amount of threat evoked on a seven-point rating scale. SDT was applied to the data by accumulating the category ratings to form a dichotomy across stimuli at each of the rating levels used to evaluate each image. Specifically, the responses in each of the rating categories to each of the six images were averaged over subjects and converted to conditional probabilities.

Other relevant studies include those in which differences in recognition memory were analyzed by sex groups (Barr-Brown and White, 1971), and the signal detection analysis of the aesthetic judgment of different

landscapes by two groups of subjects (Daniel, Wheeler, Boster, and Best, 1973).

It should be noted, however, that the collection of data summed over a group of subjects may result in values of detectability and bias which, for individual subjects, may be in considerable error. This procedure also will likely underestimate to an unknown extent the detectability index derived for each group (McNicol, 1972, pp. 111-113). These are issues which prior studies have not addressed. Hence, I performed a simulation experiment to evaluate not only (1) the effect of grouping of individual responses, but also (2) the signal prior probability and (3) the number of stimulus observations, on detectability. The details and the results of this simulation experiment are presented in Chapter 6.

## CHAPTER 6

### EFFECTS OF PRIOR SIGNAL PROBABILITY, NUMBER OF TRIALS, AND POOLING OF RESPONSES ON DETECTABILITY

As indicated earlier, signal detection theory (SDT) was first used for investigations of radar signals. Since its beginning, however, SDT has been applied to many sensory detection tasks in psychology (Banks, 1970; Green and Swets, 1966). Most of the requirements of SDT, some of which are discussed below, are satisfied in such contexts.

First, signal detection analysis is primarily applicable to experiments in which there is a large number of observations. The proportion of signal-plus-noise (SN) and noise-only (N) events in a given set of observations is usually set at 0.5 (McNicol, 1972, p. 100). The sequence of presentation of these events is assumed to be random or random appearing. Second, most of the assumptions underlying the application of SDT could easily be satisfied in such experiments. Third, these types of experiments normally assume that the subject uses a fixed

judgment criterion under a given experimental condition. To derive the ROC curve, changes in the criterion could be induced by (1) explicit instructions to the subjects to change their judgment criterion, (2) changes in the payoff functions for correct and incorrect responses, and (3) changes in the prior signal probability [i.e.,  $P(SN)$ ] (Egan and Clarke, 1966).

The basic features of signal detection analysis can, therefore, be summarized as follows: (1) a large number of trials for each subject, essentially because of the need to estimate a pair of distributions, (2) responses analyzed for each observer since detectability is normally considered a property of individuals, and (3) the proportion of SN and N events should, ideally, be equal [i.e.,  $P(SN) = P(N)$ ], unless it is varied to obtain different operating points on the ROC curve.

Recent developments, especially the extended application of SDT to other types of detection tasks where some of the usual assumptions of SDT are not satisfied, have encouraged researchers to evaluate the likely impacts of the violation of any of these assumptions [e.g.,  $P(SN) = P(N)$ ] on observers' performances. The purpose of this chapter, therefore, is to provide evidence on (1) the effects of variations in the number of trials, (2) the effects of prior signal probability, and (3) the effect on detectability of the pooling of responses across subjects.

Since in this study,  $P(SN)$  is not equal to 0.5 and there is a small number of trials which will have to be pooled, evidence obtained relating to the likely effects of these factors will provide a basis for developing expectations about the performances of the participants. Before discussing the details of a simulation analysis, herewith is a discussion of the three factors noted above.

#### Prior Signal Probability

As noted earlier, the prior signal probability  $P(SN)$  is usually set at 0.5 in signal detection experiments. However,  $P(SN)$  could be varied when the experimenter is interested in the effects of  $P(SN)$  on the observers' performances (McNicol, 1972, p. 100). In particular, such variations could be used to encourage the observer to change his/her judgment criterion in order to derive an ROC curve. SDT essentially assumes that variations in  $P(SN)$  can be used to alter an observer's judgment criterion without any change in his/her degree of sensitivity.

However, interest in evaluating the effects of variations in  $P(SN)$  arose from questions regarding the validity of the assumption that changes in  $P(SN)$  affect only the judgment criterion but not detectability. If this assumption holds, then it is possible to manipulate  $P(SN)$  to obtain a locus of points on the ROC curve each of

which represents a different judgment criterion. But, as Schulman and Greenberg (1970) indicate, only if evidence indicates that variations in  $P(SN)$  do not affect detectability is it possible to summarize the locus of points by a single fitted operating characteristic.

Previous studies on the issue of the effects of variations in  $P(SN)$  on observers' detectability have produced mixed results. For example, Nachmias (1968), Hume (1974) have reported that there was no relationship between changes in  $P(SN)$  and detectability ( $d'e$ ) or slope of ROC line(s). These studies, however, found a consistent relationship between variations in  $P(SN)$  and hit and false alarm rates. The nature of this relationship also was found to be dependent on whether or not the subjects were informed that  $P(SN)$  will vary for each experimental session. Results indicate that, with uninformed subjects, both the hit and false alarm rates decrease as  $P(SN)$  increases (Nachmias, 1968). With informed subjects, however, both the hit and false alarm rates increase as  $P(SN)$  increases (Tanner, Haller and Atkinson, 1967). In all cases, these studies conclude that observers' performances under all  $P(SN)$  conditions could be summarized by the same ROC curve.

The results reported by other studies, however, suggest a relationship between variations in  $P(SN)$  and  $d'e$ . Markowitz and Swets (1967) report that higher  $d'e$

values could be expected for higher  $P(SN)$ . In particular, their study's results indicate that  $d'e$  does increase with increasing frequency of signal presentation, at least at the higher signal-to-noise ratios. They conclude, therefore, that it does seem as if various a priori probabilities of signal presentation will yield distinct ROC curves. Also, Ogilvie and Creelman (1968) reported that for a given number of trials in an experiment, equal numbers of SN and N stimuli give the most reliable (least error variance) estimate of the points for the ROC curve.

The inconclusive results of earlier studies could be attributed to many unknown factors, including the effects of the experimental setting or the characteristics of the subjects used. For example, those studies which suggest a relationship between variations in  $P(SN)$  and  $d'e$  did not indicate whether (a) the subjects were informed of the changes in  $P(SN)$  between one experimental session and the other, or (b) if any feedback is provided to the subjects after each experiment. Nevertheless, this inconclusiveness suggests that additional evidence is needed regarding the effects of variations in  $P(SN)$  on detectability.

#### Number of Trials

As indicated earlier, detection analysis is ordinarily applicable to experiments in which a large



number of responses is elicited from a subject. The primary motive for this requirement is to enable one reliably to estimate the index of detectability and other attributes calculated from these responses.

The number of responses required for a signal detection experiment depends on the degree of reliance the researcher intends to place on the results of a study. For example, Pollack and Hsieh (1969) have, through a simulation experiment, provided an estimate of the variance of the area under the ROC curve (i.e.,  $A_c$ ). They indicate that the variance of  $A_c$ , say  $V(A_c)$ , would be consistently somewhat less than the binomial variance associated with a score on a two-alternative, forced-choice task of  $n/2$  questions; that is,

$$V(A_c) = [A_c(1-A_c)]/(n/2) \quad (6-1)$$

A somewhat conservative estimate of the sample size (number of responses) required for a 90% confidence that the estimated  $A_c$  is within plus or minus 0.05 of the true value of  $A_c$  then can be determined through the following equation:

$$2\sqrt{\frac{A(1-A)}{2/2}} = 0.05 \quad (6-2)$$

so that  $n = 2A(1-A)/(0.025)^2 = 672$  for  $A_c = 0.7$

This huge data requirement is seldom met in even the most thorough studies but the need is well recognized. For example, Green and Swets (1966) suggest that about 250 SN and 250 N events are desirable. In many applications when the stimulus material or the judgments to be made are complex, it is impossible to have as many trials as is desired to estimate individual detectabilities accurately. The alternative is then to use a number of subjects for each condition and average over subjects, a enough large number to offset the inter-subject variability. As indicated earlier, Angus and Daniel (1974) applied SDT to an experiment in marketing, in which a panel of judges rated the richness of only 27 different ice cream products on a 10-point certainty scale. It is anticipated that SDT will continue to be used in such cases with few a small number of trials. However, no previous research except Pollack and Hsien (1969) has reported which evaluates the likely effects of the number of trials (say, NT) on observers' detectability, and they used model that did not assume fixed criterion response categories. Moreover, there may be an interaction between  $P(SN)$  and NT. This also is an issue not previously addressed for the measure  $A_c$ . A preliminary simulation experiment is reported in the next section providing evidence bearing on the main and interaction effects of variations in  $P(SN)$  and number of trials on observers' performances measured by  $A_c$ .

### Pooling of Responses

Because of the problem associated with generating a sufficiently large number of trials to satisfy the requirements of SDT, some researchers (e.g., Drury and Addison, 1973; Chapman and Feather, 1971) have resorted to pooling the small number of responses across individual observers. But, as McNicol (1972) indicates, this procedure likely will underestimate to an unknown extent the detectability index derived for each group. He, therefore, advised that neither the hit and false alarm rates nor the raw data for each subject be combined. He suggested that only the  $z(s/S)$  and  $z(s/N)$  values derived from the raw data should be combined since only these will give an unbiased estimate of detectability,  $d'$ .

The responses provided by this study's subjects were pooled, in view of the following. First, each subject provided only twenty (20) responses across two experimental cases. Consequently, there will be many cells with zero observations if the responses were analyzed by individual, since the responses had to be spread over eleven response categories (see Appendix B for procedures used to analyze this study's data). However, the Grey and Morgan's (1972) program used in this study requires no cells with zero observations. Second, the number of individual responses is so small that the Beta estimates would be quite ambiguous without pooling.

It is not possible to state a priori the effects of pooling on  $A_c$ , for the following reasons. The pooled data is derived from subjects with different decision criteria and different detectabilities. If there is sample size bias, as will be indicated shortly, pooling reduces it to the extent that the subjects are similar in their response generation process. However, pooling may significantly increase the bias to the extent that individual differences, when pooled, increase the SN and N variance. The more dominant of these effects will, therefore, have to be determined empirically for each situation.

Although empirical evidence regarding the effect of pooling of the responses by this study's subjects is provided in Chapter 7, a thorough investigation of the effect of pooling of responses on  $A_c$  is left for future research.

#### Simulation Experiment

I performed a simulation experiment to provide evidence bearing on the main and interaction effects of prior signal probability  $[P(SN)]$  and the number of observations on  $A_c$ . I used  $A_c$  as an index of detectability in this simulation because (1) as indicated earlier, it is a distribution-free measure of detectability (Green 1964) and (2) it allows for a direct

comparison of the results of the simulation to the performance of the subjects in the actual experiment.

To generate the SN and N observations for each cell, I assumed the population index of detectability,  $d'$ , to be  $\sqrt{2}$ . Then, I used the cutoff values that would make the subjective probability response set  $\{.1, .3, .5, .7, .9\}$  perfectly calibrated [reported in Ferrell and McGoey (1980)] to partition into cells the observations generated. For simplicity, the probability density function  $f(y/T)$  of the decision variable  $Y$  when the proposition is true, and  $f(y/F)$  when the proposition is false, are assumed to be normally distributed with means  $d'/2$  and  $-d'/2$  respectively and with unit variance (see Figure 4-6 for a graphical representation of the model's assumptions). A pseudo random number generator produced responses according to this conventional signal detection model.

To provide evidence regarding the effects of  $P(SN)$  and number of trials on  $A_c$ , I performed a two-way analysis of variance (ANOVA). The number of observations used for the simulation ranged from 50 to 1000 (i.e., 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) while the  $P(SN)$  set used ranged from .1 to .9 (i.e., .1, .2, .3, .4, .5, .6, .7, .8, .9). Each trial and  $P(SN)$  pair was replicated four times, thus generating four  $A_c$  values per pair. I used the computer program written by Grey and

Morgan (1972) to analyze the data.

Also, to provide a basis for evaluating the impacts of  $P(SN)$  and number of observations on the performance of this study's subjects, I generated another set of  $Ac$  values from a set of observations ranging from 20 to 1000 (i.e., [20,50,100,500,1000] and a  $P(SN)$  set of [.4,.5]. These values of  $P(SN)$  correspond with those in the first account classification category used in the actual experiment (see Figure 5-2 for details). The observation set includes a sample size of twenty (20), which also corresponds with the number of responses elicited from each subject in the actual experiment.

#### Analysis and Discussion of Results

The 44x9 matrix of  $Ac$  values generated in the first simulation experiment is shown in Table 6-1. This data matrix was used to run a two-way ANOVA test of the following hypotheses:

H01:  $P(SN)$  has no significant effect on  $Ac$   
Ha1:  $P(SN)$  has a significant effect on  $Ac$

H02: The number of trials ( $N$ ) does not significantly affect the value of  $Ac$

Ha2: The number of trials ( $N$ ) does significantly affect the value of  $Ac$

The results, which are shown in Table 6-2, indicate the significance of both the main and the interaction effects

Table 6-1. Simulated Values of Area  
Under the ROC Curve

50	.66444 <sup>1</sup>	.68750 <sup>2</sup>	.78130 <sup>3</sup>	.78333 <sup>4</sup>	.82640 <sup>5</sup>	.78667 <sup>6</sup>	.79596 <sup>7</sup>	.70125 <sup>8</sup>	.73106 <sup>9</sup>
100	.69389	.79750	.76595	.84500	.74900	.83625	.77256	.75698	.65335
200	.83792	.75055	.77095	.78531	.82340	.75401	.77792	.81375	.82751
300	.73660	.82809	.81042	.80609	.83173	.83231	.83879	.81951	.80153
400	.81958	.81719	.85122	.83229	.82421	.82898	.83437	.78758	.87231
500	.80304	.81270	.81071	.81593	.80971	.83418	.84237	.81078	.82111
600	.85883	.82988	.80239	.82323	.82703	.81375	.84524	.83837	.76793
700	.83110	.83857	.82882	.82187	.81561	.83287	.83495	.81217	.80372
800	.83818	.82805	.80346	.80224	.82717	.79414	.80601	.81542	.82525
900	.82427	.83140	.84396	.81082	.81601	.82301	.81597	.84202	.81172
1000	.80461	.83307	.80315	.82986	.82038	.84308	.81785	.83218	.83672
50	.65111	.74625	.76476	.73583	.85040	.77583	.63787	.72125	.69647
100	.62055	.76469	.74119	.82625	.84120	.74104	.75643	.79406	.75507
200	.68611	.79836	.78055	.81844	.84675	.81693	.74832	.84773	.77747
300	.79401	.83892	.78098	.84255	.84262	.83438	.83101	.83913	.81539
400	.87063	.78730	.80750	.79307	.81954	.82111	.76700	.82969	.73156
500	.81027	.81520	.84857	.83495	.83377	.82213	.79277	.82065	.77228
600	.74909	.83411	.82657	.80432	.82602	.83165	.76632	.82043	.82256
700	.81924	.77903	.83215	.80065	.82195	.82538	.83387	.83034	.83754
800	.81781	.80384	.83031	.83220	.80691	.83050	.82162	.83967	.78374
900	.75399	.84348	.81827	.83054	.81249	.82410	.83099	.83080	.74826
1000	.78983	.82941	.79874	.84226	.81636	.81800	.85263	.84097	.84451
50	.63333	.81000	.86381	.83000	.79520	.80333	.70680	.68750	.69886
100	.72556	.76563	.74024	.80583	.81420	.81813	.73048	.79781	.79573
200	.74847	.82859	.82857	.74245	.79090	.82167	.77332	.78781	.76230
300	.78444	.80007	.80053	.80678	.80791	.80370	.77315	.84698	.71503
400	.83969	.80943	.80939	.82243	.83641	.76948	.79343	.84590	.84337
500	.82820	.84274	.84263	.81184	.81962	.85437	.81152	.82691	.79811
600	.79073	.81232	.79972	.81286	.87877	.79116	.77582	.79417	.80532
700	.82474	.84285	.82655	.83356	.85216	.80418	.85033	.81739	.83972
800	.81781	.80171	.81782	.81358	.82915	.79310	.82468	.82821	.81330
900	.82755	.81922	.84179	.82273	.81883	.80594	.84534	.80736	.83554
1000	.84617	.82074	.84311	.82108	.82993	.81419	.84688	.83259	.80510
50	.67333	.77750	.60571	.80000	.78640	.67083	.70680	.72250	.64394
100	.75222	.84094	.85976	.84896	.79880	.81271	.83661	.84656	.76813
200	.78028	.81453	.76839	.83563	.78500	.84594	.80570	.80031	.72918
300	.80735	.83177	.82921	.78757	.81536	.79275	.85225	.82174	.76546
400	.72458	.78420	.81250	.80660	.82366	.81572	.81260	.83352	.79594
500	.83647	.81600	.82987	.80865	.80783	.82139	.79726	.80306	.81482
600	.75981	.78296	.81637	.82262	.84227	.83914	.80322	.83398	.81272
700	.82973	.83154	.85533	.83025	.84498	.83361	.84309	.82031	.83985
800	.82797	.80172	.82754	.79168	.81603	.81786	.85168	.83732	.80520
900	.80179	.80195	.82493	.84345	.83459	.84398	.79538	.81650	.82033
1000	.84117	.82313	.83016	.81311	.83695	.82526	.81373	.81828	.82973

**Table 6-2. Two-Way Analysis of Variance: Effects of Number of Trials (NT) and Prior Signal Probability P(SN) on Area Under the Curve.**

<b>Main Effects</b>	.303	18	.017	18.71	.001
N	.057	8	.007	7.90	.001
P(SN)	.246	10	.025	27.36	.001
<b>2-Way Interactions</b>					
(N & P(SN))	.122	80	.002	1.70	.001



of  $P(SN)$  and number of observations on  $Ac$ . The observed significance of the effect of  $P(SN)$  on  $Ac$  agrees with the results of the study reported by Markowitz and Swets (1967), in which they experimentally found that different  $P(SN)$  may yield distinct ROC curves. However, the mechanism may be entirely different in the two cases.

An overview of the results indicates that the  $Acs$  calculated for small  $NT$  are generally less than those for large  $NT$ . Furthermore,  $Acs$  for low  $P(SN)$  have a higher degree of variability (i.e., have a wider range of values) than those of higher  $P(SN)$  in accord with Ogilvie and Creelman's (1968) observations that, for a given number of observations in an experiment,  $P(SN) = P(K) = 0.5$  gives the most reliable estimate of the points for the ROC curve. Relative to the expected  $Ac$  of 0.54 for the population from which the sample observations are assumed to have come, there seems to be a consistent underestimation of  $Ac$ . The degree of underestimation is lower for large  $NT$ , although the effect of this (sample size) on the degree of underestimation is reduced as  $P(SN)$  approaches 0.5. This finding suggests an interaction effect of  $P(SN)$  and  $NT$ , as indicated earlier. This result is presented graphically in Figure 6-1.

To provide further evidence regarding the effects of variations in  $P(SN)$  and the number of observations (or responses) on the performance of this study's subjects, I

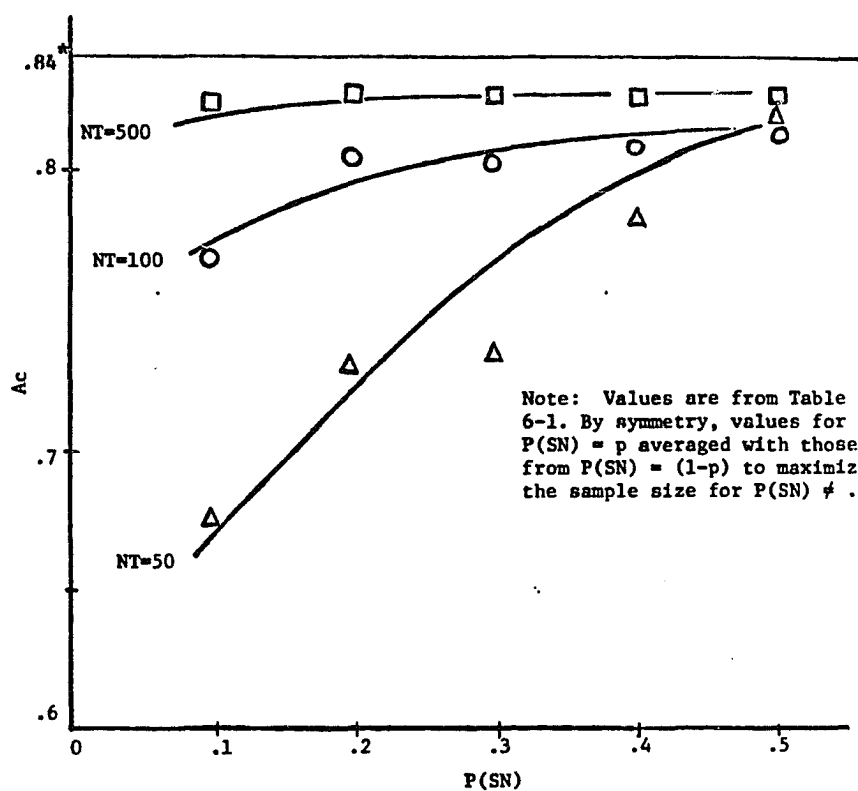


Fig. 6-1. Effect of  $P(SN)$  and Number of Observations ( $NT$ ) on the Underestimation of Area Under the ROC Curve.

\* Expected  $Ac$  for the population with a detectability of  $\sqrt{2}$ .

generated two hundred and twenty-five (225)  $A_c$  values according to the above simulation model for each given number of trials ( $N$ ) for values of prior signal probability  $P(SN)$  of 0.4 and 0.5. The results are shown in Table 6-3.

#### Comment on the Simulation Results

In the simulation, the signal detection model was assumed and pseudo random numbers generated according to that model were used as responses to calculate  $A_c$  as a function of  $P(SN)$  and number of trials. The results indicate an underestimation bias which is greater with smaller  $P(SN)$  and number of observations. Such bias was not found by Pollack and Hsieh (1969) and has not been reported elsewhere, and no clear theoretical explanation for it has yet emerged.

Moreover, the standard deviations for  $A_c$  are a little more than half the corresponding binomial standard deviation and in the case of the estimate most closely corresponding to the conditions of Pollack and Hsieh (1969) simulation (i.e., 100 total samples and  $P(SN) = 0.5$ ), they obtained a substantially larger standard deviation, 0.055, as compared with 0.034 in Table 6-3. Their value is based on 100 independent simulations and that in the table are over 200, so the effects is not due to small sample bias in estimation of variance.

Table 6-3. Effects of Number of Trials (NT) and Prior Signal Probability P(SN) on Variability of Area Under the Curve

		A R E A   U N D E R   T H E   C U R V E						
N U M B E R   O F   T R I A L S		R A N G E		Mean	Standard Deviation	Binomial Standard Deviation	Coefficient of Variation	
		Lowest	Highest					
20	.4	.402	.782	.614	.066		.108	
	.5	.442	.740	.646	.063	.15	.097	
50	.4	.593	.810	.730	.047		.065	
	.5	.600	.832	.743	.050	.09	.067	
100	.4	.668	.840	.773	.035		.045	
	.5	.692	.842	.780	.034	.06	.044	
500	.4	.754	.835	.803	.018		.022	
	.5	.759	.837	.805	.018	.03	.022	
1000	.4	.761	.836	.803	.014		.016	
	.5	.773	.833	.809	.013	.02	.016	

Although much effort and care has been put into checking the simulation program and the method of calculating  $A_c$ , it is still possible that the simulation is faulty. Until the bias and variance properties of  $A_c$  are confirmed by an independent simulation that uses different procedures but the same underlying model, these results must be considered tentative. If correct, their implications are quite serious, so further research on this matter is imperative.

It is concluded that with the number of responses per auditor (20) and the number of auditor-subjects (28) both as large as it was feasible to make them, it is suitable to pool the results in order to calculate  $A_c$  and decision criterion (Beta). If the measure  $A_c$  is biased as the simulation suggests, pooling will reduce the bias. Additionally, there can be expected underestimation due to the pooling of different criteria and detectability, but it will be slight.  $P(SN)$  has been made as near 0.5 as feasible and, even if the simulation is correct, the effect of  $P(SN) = 0.4$  on the pooled data should not produce a very substantial underestimation.

## CHAPTER 7

### THE DATA ANALYSIS

In this chapter, I present an analysis of the subjects' responses in accordance with the research issues identified in Chapter 5.

#### Discussion of Results by Research Issue :

Research Issue Number One:

Detectability of Auditors' Responses

To provide evidence relating to this research issue, a computer program based on the maximum likelihood estimation model suggested by Grey and Morgan (1972) was used to derive the index of detectability of auditors' responses. For reasons stated earlier (see Chapters 4 and 5),  $A_c$ , which is equivalent to the area under the ROC curve, was the chosen measure of the subjects' detectability.

Table 7-1 shows the  $A_c$ s derived from these responses across all experimental conditions. An overview of the results indicates that, overall, the  $A_c$ s range from 0.574 to 0.714 across the three account classification approaches.

Table 7-1. Area Under the ROC Curve

Level of Analysis	Item	Account Classification		
		Approach		
		I	II	III
Overall		.648	.574	.714
Internal Control	ABC	.673	.482	.751
	XYZ	.616	.679	.673
Functional Level	Manager	.716	.587	.782
	Senior	.620	.567	.689

Of interest also is the effect of the experimental variables on Ac. Table 7-1 indicates that, in general, the Acs attained when the internal control system is adjudged strong are higher than the Acs attained when the internal control is adjudged weak. This finding suggests that auditors are likely to make more accurate judgments when they face a stronger AIC environment than when they face a weaker AIC environment. A discussion of the effects of the state of internal control on the characteristics of the subjects' responses is provided below.

An evaluation of Ac by functional level also indicates that, for all account classification approaches, audit managers' Acs are greater than that of the audit seniors. This finding is consistent with the expectation that the more experienced audit managers' judgmental accuracy should be higher than those of the less experienced audit seniors. An evaluation of the differences in the characteristics of the judgments by functional level accounting for these differences in Acs will be presented shortly.

In general, the auditors' Acs range from 0.482 to 0.782. This should not be unexpected, however, given the dampening effects of, say, the small number of responses per subject on Ac noted in Chapter 6. Relative to the results of the simulation experiment reported in Chapter



6, the Acs suggest a reasonably high degree of detectability by the auditor-subjects.

Consistent with the findings reported for the simulation experiment, there seems to be an effect of  $P(SN)$  on the variability of the subjects' Acs. The Acs reported for the higher  $P(SN)$  in respect of the first account classification approach are less variable (they range from 0.616 to 0.716) than those of the second and the third account classification approaches (which range from 0.482 to 0.714). As indicated in Chapter 6, it is impossible to state a priori the effect of pooling of responses on Ac. To provide preliminary evidence on the effect of pooling of responses on this study's subjects, I calculated the Acs by individual subject for the first account classification category, which were then averaged for each level of analysis. The Acs calculated both by averaging and pooling are reported in Table 7-2. The results suggest that in general, pooling causes an underestimation of Acs, although this effect appears insignificant. However, as indicated in Chapter 6, further research on this matter is suggested.

#### Research Issue Number Two:

##### Subjects' Judgment Bias

Evidence bearing on this research issue indicates the nature of the subjects' preferences for one category

**Table 7-2. Effect of Pooling of Responses on Area Under the ROC Curve**

	Procedure	
	Averaged Acs	Pooling of Responses
Overall	0.660	0.648
Internal Control:		
ABC	0.675	0.673
XYZ	0.646	0.646
Functional Level		
Manager	0.703	0.716
Senior	0.643	0.620

of response (say, "sn") to another (say, "n"). It also allows an evaluation of the effects of the account categorization approaches on the subjects' decision rules.

As shown in Chapter 4 and in Chapter 5, the value of the ratio of the regret for incorrectly responding "sn" to the regret of incorrectly responding "n" (i.e.,  $K$ ) provides a measure of the subjects' response bias. A value of  $K$  greater than one indicates that the subjects are more prone to responding "n", while  $K < 1$  suggests that the subjects were more inclined to respond "sn". A  $K$  value equal to one ( $K = 1$ ) indicates indifference on the part of the subjects.

Table 7-3 presents the derived measure of bias (Beta) and the associated  $K$  values for each account classification approach. The  $K$  values indicate that, overall, the auditor-subjects were not indifferent to the costs of decision errors they are likely to commit. The  $K$  value of 0.89 reported suggests that the auditors were more prone to respond "sn" than "n". This finding suggests that, in general, auditors tend to avoid, or at least minimize, the costs of incorrect acceptance of materially misstated account book values.

The AIC environment appears to have the most pronounced effect on the nature of the subjects' judgment bias. When the internal control is adjudged strong, the subjects were more prone to respond "n" ( $K = 1.19$ ) than

Table 7-3. Index of Response Bias by Account Classification Approach

Level of Analysis	Item	I		II		III	
		Beta	K	Beta	K	Beta	K
Overall		1.09	.89	1.15	.29	.967	.24
Internal Control	ABC	1.45	1.19	1.64	.41	1.13	.28
	XYZ	.89	.69	.93	.23	.92	.23
Functional Level	Manager	1.04	.85	1.23	.31	..86	.21
	Senior	1.09	.89	1.12	.28	..99	.25

"sn". However, when the internal control is perceived to be relatively weak, the subjects were more prone to respond "sn" ( $K = 0.69$ ) than "n". This finding suggests that the subjects were generally more skeptical of the fair presentation of the account book values when facing a weaker AIC environment. Similarly, they were more confident of the fair presentation of the account book values when facing a stronger AIC environment. As a result, consistent with professional standards, auditors would prefer to perform more tests of details when the AIC system is adjudged weak than when the AIC system is perceived to be strong.

Functional level does not seem to have a significant effect on the subjects' decision rules. Both the managers and the seniors displayed a tendency towards responding "sn" rather "n". This finding suggests that both groups of auditors are equally disposed to minimizing the costs associated with errors of incorrect acceptance of materially misstated book values.

The  $K$  values reported under the second and the third account classification approaches should, however, be interpreted with caution, given the potential confounding effects of  $P(SN)$  on these  $K$  values. For these two approaches,  $P(SN)$  is 0.2. From equation (4-17), it should be apparent to the reader that a low  $P(SN)$  will have a dampening effect upon the value of  $K$ . Therefore,

since the low K values reported in these situations are confounded with the effect of the low P(SN) values, no meaningful inferences about the subjects' preferences could be made under these two situations.

An overview of both the Beta and the K values, however, suggest a relationship similar to those observed under the first classification approach. For example, the differences in K values are greatest for the AIC variable, but they are almost imperceptible for the functional level variable. In relative terms, the K values for the weak AIC environment are greater than the K values for the weak AIC environment. Therefore, the observations made about the subjects' preferences under the first account classification approach appear applicable to those of the second and the third account classification approaches.

#### Research Issue Number Three:

##### Effect of Task Criterion on Auditors' Performance

As indicated earlier, Tukey's (1960) distinction between conclusions and decisions suggest that the responses provided by this study's subjects should be affected by the task criterion. In particular, the subjects are expected to be more sensitive to the consequences of their decision errors under the action criterion than under the judgment criterion. To provide

evidence bearing on this idea, I evaluated the significance of the differences in the subjects' responses by task criterion in terms of (a) proportion of correct responses, (b) false alarm rates, and (c) miss rates. I also compared the subjects' index of bias by task criterion.

A priori, one can state that, if the subjects are more concerned with the consequences of their decision errors under the action criterion, then their decision rule under the action criterion should be more stringent than their decision rule under the judgment criterion. That is, the auditors will be more prone to respond "sn" than "n" under the action criterion.

This scenario also implies that, when making actual audit decisions, the regret which auditors associate with false alarms should be less than those associated with misses. Hence, the K values for the action criterion should be lower than those of the judgment criterion. I provide evidence bearing on this idea through a comparison of the auditors' judgmental biases (Beta) and regret ratio (K values) by task criterion.

Table 7-4 presents a summary of these characteristics for each classification approach by task criterion. To determine the effect of task criterion with respect to a given classification approach, the test for

Table 7-4. Characteristics of Auditor's Responses

## A: Account Classification I

(i) Judgment

		RESPONSE	
		sn	n
STATE OF NATURE	SN	140	112
	N	97	211

(ii) Action

		RESPONSE	
		sn	n
STATE OF NATURE	SN	147	109
	N	135	169

## B: Account Classification II

(i) Judgment

		RESPONSE	
		sn	n
STATE OF NATURE	SN	57	55
	N	180	268

(ii) Action

		RESPONSE	
		sn	n
STATE OF NATURE	SN	74	38
	N	223	225

## C: Account Classification III

(i) Judgment

		RESPONSE	
		sn	n
STATE OF NATURE	SN	76	36
	N	157	291

(ii) Action

		RESPONSE	
		sn	n
STATE OF NATURE	SN	73	39
	N	221	227



differences in proportion for nonindependent samples suggested by some researchers (for example, McNemar [1949]; Cochran [1950]) was used to analyze the data. The samples were considered nonindependent since the same set of subjects provided responses twice for each account item, since the experimental task requests the subjects to provide a second response (that is, to state whether an account item would require special audit attention) in light of their earlier responses. The statistical test enables one to evaluate the significance of the changes in the characteristics of the subjects' responses (e.g., from being "correct" to being "incorrect") due to a change in the task criterion. An example will make this clear.

In Table 7-5, I present a summary of the changes in the number of correct responses for each subject with respect to the account items under the first classification category. For example, 14 of the first subject's responses under the judgment criterion were correct, while only 13 of these responses were correct under the action criterion. The net effect of the change in response criterion, therefore, is to reduce by one the number of this subject's responses under the action criterion. Similarly, with respect to the tenth subject, the net effect of the change in response criterion is to increase by two the number of correct responses under the action criterion. The totals of these changes (43 and 8

Table 7-5. An Example of the Procedure for Evaluating the Effect of Task Criterion on the Subject's Responses.

Item: Proportion of Correct Responses With Respect To Account Classification Category I

Subject	Correct Responses		Net Difference	
	Judgment	Action	Judgment	Action
1	14	13	1	--
2	14	9	5	--
3	17	15	2	--
4	14	12	2	--
5	11	8	3	--
6	12	12	--	--
7	12	14	--	2
8	14	9	5	--
9	11	12	--	1
10	11	13	--	2
11	9	9	--	--
12	16	17	--	1
13	13	11	2	--
14	15	10	5	--
15	11	11	--	--
16	12	13	--	1
17	13	11	2	--
18	11	9	2	11
19	11	11	--	--
20	9	9	--	--
21	11	10	1	--
22	13	12	1	--
23	11	12	--	1
24	11	10	1	--
25	14	12	2	--
26	13	10	3	--
27	13	12	1	--
28	15	10	5	--
Total	351	316	43	8

for judgment and action criterion respectively) are also shown. It is these totals which are used to evaluate the significance of the differences in the subjects' responses due to a change in the response criterion.

This scenario has been represented graphically by Glass and Stanley (1970) and could be adapted for evaluating this study's data as follows:

J U D G M E N T		
	Correct	Incorrect
Incorrect	A	B
Correct	C	D

The test for differences in proportions for nonindependent samples focuses only on A and D, which are the number of responses that changed with respect to a defined attribute from one situation to another. For example, from Figure 7-5, the total of 43 net responses for the judgment criterion is analogous to A, while the total of 8 net responses under the action criterion is analogous to D. Note also that the test statistic does not require a calculation of the proportions  $A/(A+B)$  or  $D/(C+D)$ ; the only values required are A and B.

Glass and Stanley (1970) indicate that the appropriate test statistic to be used to test the null hypothesis (that is, that there is no significant

difference in the proportion of the responses) against the alternative hypothesis can be represented as follows:

$$z = \frac{D - A}{\sqrt{(D + A)}} \quad (7-1)$$

where  $z$  is normally distributed with a mean of zero and variance of 1 if and only if  $(D + A)$  is greater than 10.

Table 7-6 presents the results of the test for (a) judgmental accuracy (that is, the total number of correct responses), (b) false alarms, and (c) misses for each experimental variable. The numbers under the headings A and D represent the total net responses under the judgment and action criteria respectively, calculated as shown in the Figure 7-4 by attribute (i.e., correct responses, false alarms, and misses). There appears to be a significant effect of task criterion on the subjects' judgmental accuracy, with Table 7-4 indicating that this significant difference could be attributed mainly to the substantial increase in the subjects' false alarms under the action criterion relative to their false alarms under the judgment criterion.

Note that, except for the second classification approach, task criterion has no effect on the magnitude of the changes in the subjects' misses. This finding may be attributed to the fact that auditors generally tend to minimize the likelihood of erroneously accepting materially misstated balances, a strategy which results in

Table 7-6. Test for Differences in Characteristics of Auditors' Responses by Task Criterion

Account Classification Approach	Item	A <sup>+</sup>	D <sup>+</sup>	Z
I	Correct Responses	45	8	-4.90 <sup>*</sup>
	False Alarms	5	45	5.66 <sup>*</sup>
	Misses	21	16	-0.82
II	Correct Responses	44	18	-3.30 <sup>*</sup>
	False Alarms	12	55	5.25 <sup>*</sup>
	Misses	20	3	-3.54 <sup>*</sup>
III	Correct Responses	77	10	-7.18 <sup>*</sup>
	False Alarms	10	74	6.98 <sup>*</sup>
	Misses	6	9	0.77

\* Significant at the 99% confidence level.

+ See Table 7-4 for the procedure used to derive these values.

low miss rates regardless of the task criterion. This finding also corroborates those findings reported under the second research issue which indicate that, in general, auditors tend to minimize the regret of incorrectly saying "sn" more than the regret of incorrectly saying "n".

In summary, it appears that the task criterion has a significant effect on the subjects' performance. It also appears that the main cause of this significant difference is that the subjects were more concerned with minimizing the error of incorrectly accepting materially misstated book values than the error of incorrectly rejecting fairly presented book values.

Discussions with many of the participants after each experiment lend support to the findings reported above. They indicated that, in practice, there are account items which always are examined in detail because of their nature and/or perceived importance, even when the reported book value conforms with the auditor's expectations. These account items, they indicate, typically include sales, accounts receivables, inventory and fixed assets. It appears, therefore, that when making audit program decisions, auditors prefer to be safe rather than be sorry. It is this strategy which, as shown above, accounts for the higher incidence of false alarm errors committed under the action criterion.

I present in Table 7-7 a comparison of the Beta and the corresponding K values by account classification approach as a basis for evaluating the stringency of auditors' decision rules under each task criterion. As hypothesized, the results indicate that auditors are more prone to responding "sn" than "n" under the action criterion. This finding is consistent with the idea that, when making actual audit decisions, auditors attach greater importance to the regret of incorrectly accepting materially misstated book values than the regret associated with incorrect rejection of fairly presented book values.

The quality of AIC also seems to have a mitigating effect on the stringency of the subjects' biases. In all cases, the subjects' biases were less severe when the AIC was adjudged strong. That is, auditors are less prone to slating for intensive audit account balances which are fairly presented when the AIC was adjudged strong than when it was adjudged weak.

In general, the auditors' decision rules seem insensitive to the prior signal probability (i.e., the base rate). This is not unexpected, since the subjects were not informed of this at any stage of the experiment. Future research should, however, endeavor to investigate the impact of base rates on the characteristics of auditors' PAR judgments.

Table 7-7. Effect of Task Criterion on Judgment Bias

Account Classifi- cation Approach	Level of Analysis	B E T A		K	
		Judgment	Action	Judgment	Action
I	Overall	1.09	.64	.89	.53
	Internal Control:				
	ABC	1.45	.57	1.19	.47
	XYZ	.84	.72	.69	.59
	Functional Level:				
	Manager	1.04	.53	.85	.44
	Senior	1.09	.70	.89	.57
II	Overall	1.15	.84	.29	.21
	Internal Control:				
	ABC	1.64	1.03	.41	.26
	XYZ	.93	.70	.23	.17
	Functional Level:				
	Manager	1.23	.85	.31	.21
	Senior	1.12	.85	.28	.21
III	Overall	.97	.57	.24	.14
	Internal Control:				
	ABC	1.13	.44	.28	.11
	XYZ	.92	.65	.23	.16
	Functional Level:				
	Manager	.86	.52	.21	.13
	Senior	.99	.59	.25	.15



The potential implications of this finding are twofold. First, it suggests that auditors will be less efficient at performing the audit task, given the inherent desire to minimize the error of incorrect acceptance of materially misstated book values. Second, the finding suggests that auditors' beliefs may not be independent of their preferences, as postulated by Bayesian theory. The implications of this finding regarding the application of the Bayesian model to the audit decision process is left for further research.

#### Research Issue Number Four:

##### The Effect of Internal Control on Auditors' Responses

To provide evidence bearing on this research issue, the test for differences in proportions for independent samples was used to analyze the subjects' responses by state of internal control for each task criterion and account classification approach. As in the third research issue, the characteristics of interest are the subjects' judgmental accuracy and decision errors. Unlike the third research issue, however, the subjects' responses were regarded as having been obtained from independent samples. As stated earlier, the experimental cases were based on data from two independent audit clients. Therefore, there is no basis for concluding that

the subjects' responses for each case were dependent.

The test statistic for differences in proportions for independent samples suggested by Glass and Stanley (1970, p.325) was used to analyze the data. The test statistic can be described as follows. Assume that there are two populations,  $N_1$  and  $N_2$ , from which the samples  $n_1$  and  $n_2$ , respectively, were taken. The number of persons in the sample from  $n_1$  possessing the characteristic of interest is  $f_1$  such that the proportion  $p_1$  is  $f_1/n_1$ . Similarly, the respective value for  $n_2$  is  $f_2$  such that  $p_2$  is  $f_2/n_2$ . Therefrom, the following test statistic is defined:

$$z = \frac{p_1 - p_2}{\sqrt{\left(\frac{f_1 + f_2}{n_1 + n_2}\right) \left(1 - \frac{f_1 + f_2}{n_1 + n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (7-2)$$

The quantity  $(f_1 + f_2)/(n_1 + n_2)$  is the proportion of responses in both samples  $n_1$  and  $n_2$  that possess the characteristic of interest. Also,  $(f_1 + f_2)/(n_1 + n_2)$  multiplied by 1 minus the same quantity, is an estimate of the variance of the dichotomously scored variable  $X$  with mean  $P$ .

If, for both populations,  $n_1 P_1$  [or  $n_1(1-P_1)$  whichever is smaller] and  $n_2 P_2$  [or  $n_2(1-P_2)$  whichever is smaller] are greater than 5, then  $z$  in equation (7-2) has a normal distribution with mean 0 and variance 1 over repeated pairs of independent samples.

Table 7-8 presents the results of the test for differences in proportions for the relevant characteristics of the subjects' responses. The results indicate that, in general, the significance of the differences in the subjects' performance for each AIC environment is inversely related to  $P(SM)$ . For example, most of the differences were not statistically significant for the first account classification approach, whereas most of the differences were significant for the second and the third approaches.

The results also reveal a tendency to overrely on the good internal control system or to underrely on the bad internal control system. For example, the miss rates were generally higher when the internal control is adjudged strong. That is, the subjects were more prone to erroneously accepting materially misstated account items when the internal control system was adjudged good. Similarly, the subjects were more prone to slating for intensive audit account balances which are fairly presented when the AIC is adjudged weak. The implications of this behavior for both the efficiency and the effectiveness of the audit have been noted earlier. That is, overreliance on good internal control negatively affects the effectiveness of the audit, while an underreliance on weak internal control will adversely affect the efficiency of the audit since, in the latter

Table 7-8. Effect of Quality of Internal Control on Auditor's Responses

Account Classifi- cation Approach		Item	C R I T E R I O N									
			JUDGMENT				ACTION					
			Internal Control			z	Internal Control			z		
A	B	C	X	Y	Z		A	B	C		X	Y
I	Judgment Accuracy**	.64			.62		.50		.60		.54	1.43
	False Alarm Rate <sup>+</sup>	.43			.50		-1.32		.55		.57	-0.48
	Miss Rate <sup>++</sup>	.52			.39		-2.06*		.45		.39	0.96
II	Judgment Accuracy	.57			.59		-0.54		.58		.49	3.02*
	False Alarm Rate	.35			.45		-3.05*		.42		.58	-4.79*
	Miss Rate	.71			.27		6.59*		.43		.25	2.84*
III	Judgment Accuracy	.73			.58		3.73*		.59		.48	2.61*
	False Alarm Rate	.24			.46		-6.90*		.40		.58	-5.39*
	Miss Rate	.36			.29		1.12		.45		.25	3.14*

\* Significant at the 95% confidence level.

\*\* Number of Correct Responses as a Percentage of Total Responses.

+ From Figure 4-2, Miss Rate =  $f_2 / (f_2 + f_4)$ ;

False Alarm Rate =  $f_3 / (f_3 + f_4)$ .

situation, the auditor will perform more tests of details than the situation warrants.

Overall, it appears that the subjects were able to make more accurate judgments when the internal control system is adjudged strong than when the system is adjudged weak. For example, the subjects' judgment accuracy was higher for the strong internal control situation than for the weak internal control situation. Also, there appears to be a slight evidence of the effect of task criterion on the subjects' judgment accuracy. For example, other than for the second classification approach, the proportion of correct responses (decisions) were generally higher (lower) under the judgment (action) criterion.

#### Research Issue Number Five:

##### The Effect of Functional Level On Auditors' Performance

As indicated earlier, SDT assumes that detectability is positively related to an observer's accumulated relevant experience in the subject matter of a detection task. Furthermore, the structure of the AR process in practice suggests that the more experienced auditors should be able to make more accurate judgments. To provide evidence bearing on this idea, the subjects' responses were evaluated by functional level. For reasons similar to those stated for the fourth research issue, the

test for differences in proportions for independent samples was used to analyze the data, since the responses by, say, the managers are in no way dependent upon the seniors' responses.

Table 7-9 presents the results of the test for the effect of functional level on the subjects' performances. The table indicates that, other than miss rates (that is, the error of incorrectly accepting materially misstated book values), the audit managers' performances were generally not statistically different from those of the seniors. That is, at least for the first and the third classification approaches, it appears that the seniors' decision errors of incorrectly accepting materially misstated account balances were more severe than those committed by the managers. This finding suggests that the managers may be more effective than the seniors at making PAF judgments.

Despite the statistical insignificance of the differences in other aspects of the subjects' responses, an overview of Table 7-9 suggests an overall superior performance by the managers. In all situations, the managers' judgmental accuracy was slightly greater than that of the seniors. Similarly, in most cases, the decision errors committed by the managers were lower than those committed by the seniors. A closer look at the results also suggests that under the judgment criterion,

Table 7-9. Effect of Functional Level on Auditors' Performance

Account Classifi- cation Approach		C R I T E R I O N					
		JUDGMENT			ACTION		
		Functional Level			Functional Level		
		Manager	Senior	z	Manager	Senior	z
I	Judgment Accuracy <sup>+</sup>	.68	.61	1.55	.61	.55	1.29
	False Alarm Rate <sup>+</sup>	.32	.31	0.17	.41	.46	-0.80
	Miss Rate <sup>+</sup>	.33	.49	-2.31 <sup>*</sup>	.36	.45	-1.31
II	Judgment Accuracy	.57	.59	-0.43	.55	.53	0.32
	False Alarm Rate	.44	.39	0.98	.48	.50	-0.38
	Miss Rate	.41	.53	-1.15	.38	.36	-0.81
III	Judgment Accuracy	.66	.66	0.00	.57	.52	1.07
	False Alarm Rate	.38	.34	0.80	.48	.50	-0.38
	Miss Rate	.19	.38	-1.94 <sup>**</sup>	.22	.40	-1.81 <sup>**</sup>

+ See notes at the bottom of Table 7-7.

\* p 0.03

\*\* p 0.08

managers appear to be less prone to committing the error of slating for intensive audit account book values which are fairly presented (i.e., greater false alarms) than the seniors. This scenario seems to persist uunder the action criterion, thus suggesting that, in practice, managers may be more capable at detecting materially misstated book values than seniors.

#### Research Issue Number Six:

##### Calibration of Subjects' Responses

As stated earlier, the aim of this research issue is to provide evidence relating to the auditors' effectiveness at communicating their knowledge or, equivalently, the extent to which they are sensitive to their level of uncertainty in the accuracy of their responses. Calibration is considered the appropriate measure for providing evidence bearing on this feature, as discussed in Chapter 4. Calibration is measured by matching the subjects' proportion of correct responses for each response category  $[P(C/r_i)]$  with the given response category  $r_i$  (that is, each subjective probability value on a scale  $[\cdot 5, 1]$ ).

Table 7-10 summarizes the relationship between the subjects' proportion of correct responses (along each row) against each subjective probability value (i.e., response category)  $r_i$  for all account classification approaches.



Table 7-10. Calibration of Subjects' Responses\*

ACCOUNT CLASSIFICATION APPROACH																			
I							II							III					
Subjective Probability							Subjective Probability							Subjective Probability					
	.5	.6	.7	.8	.9	1.0	.5	.6	.7	.8	.9	1.0	.5	.6	.7	.8	.9	1.0	
Overall	.7	.5	.6	.7	.7	.8	.6	.6	.6	.6	.6	.4	.6	.6	.6	.7	.8	.6	
ABC	.7	.5	.5	.7	.8	.8	.7	.5	.5	.6	.6	0.0	.7	.6	.7	.7	.9	.8	
XYZ	.7	.5	.6	.7	.6	.9	.4	.6	.6	.7	.6	.6	.4	.6	.6	.7	.7	.6	
Manager	.8	.5	.6	.7	.7	1.0	.7	.5	.5	.6	.6	.6	.6	.5	.6	.7	.8	.8	
Senior	.6	.5	.6	.6	.7	.7	.5	.6	.6	.7	.6	.2	.6	.6	.6	.7	.7	.5	

\* Figures along the row represent the subjects' proportion of correct responses per given subjective probability.

To facilitate an interpretation of this summary, the results are reproduced in graphical form in Figures 7-1 to 7-7.

Figure 7-1 indicates that, as one might expect, the subjects' responses were less than perfectly calibrated. The figure also indicates that overconfidence is the predominant nature of the miscalibration, except for the judgment criterion at the 0.5 response category, which suggests a tendency towards underconfidence.

Figures 7-2 to 7-4 present the calibration of subjects' responses by state of internal control for the three account classification approaches. The figures indicate that the nature of miscalibration is, also, one of overconfidence, except for the responses at the 0.5 response category for the judgment criterion, which indicate a tendency towards underconfidence. There does not seem to be a significant effect of the state of internal control on the calibration of the responses.

Finally, Figures 7-5 to 7-7 show, for each account classification approach, the effect of functional level on the calibration of the subjects' responses. The pattern of miscalibration is identical to those described above, that is, predominant overconfidence. Also, it appears there is no significant effect of functional level on the calibration of these responses.

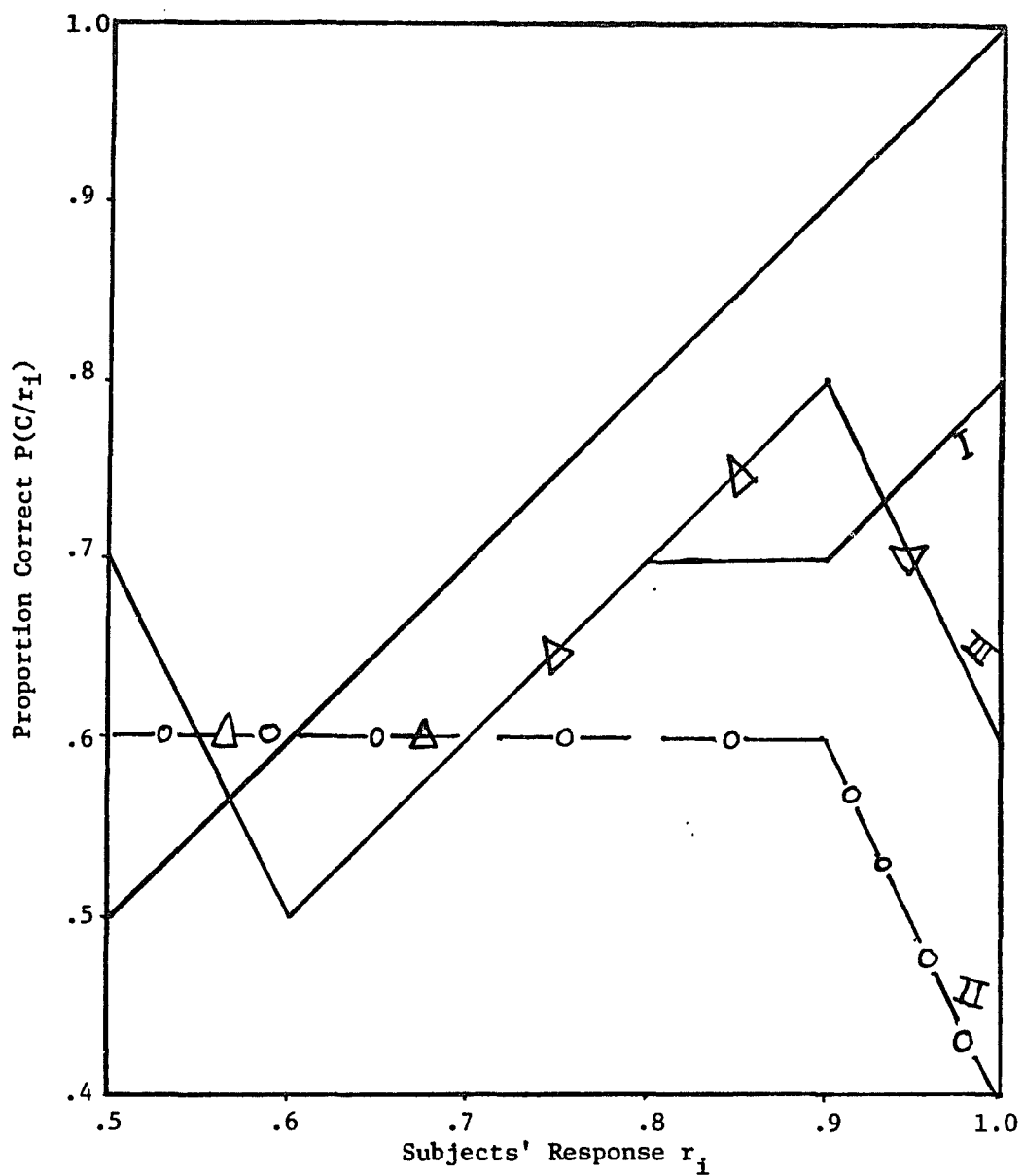


Fig. 7-1. Overall Calibration of Subjects' Responses by Account Classification Criterion.

— I = Account Classification Criterion I  
 —○— II = Account Classification Criterion II  
 —△— III = Account Classification Criterion III

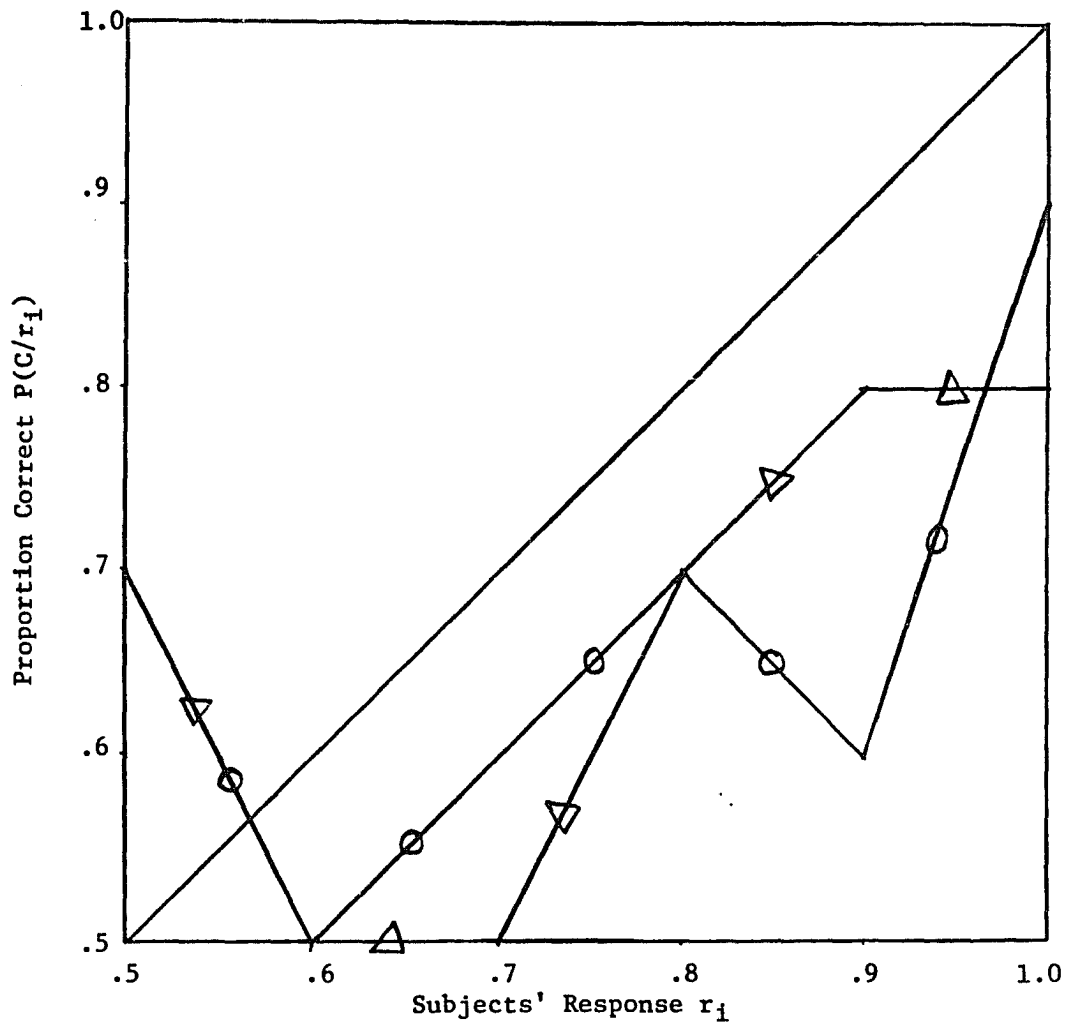


Fig. 7-2. Account Classification Criterion I: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses.

$\triangle$  = Good AIC System (ABC)  
 $\circ$  = Bad AIC System (XYZ)

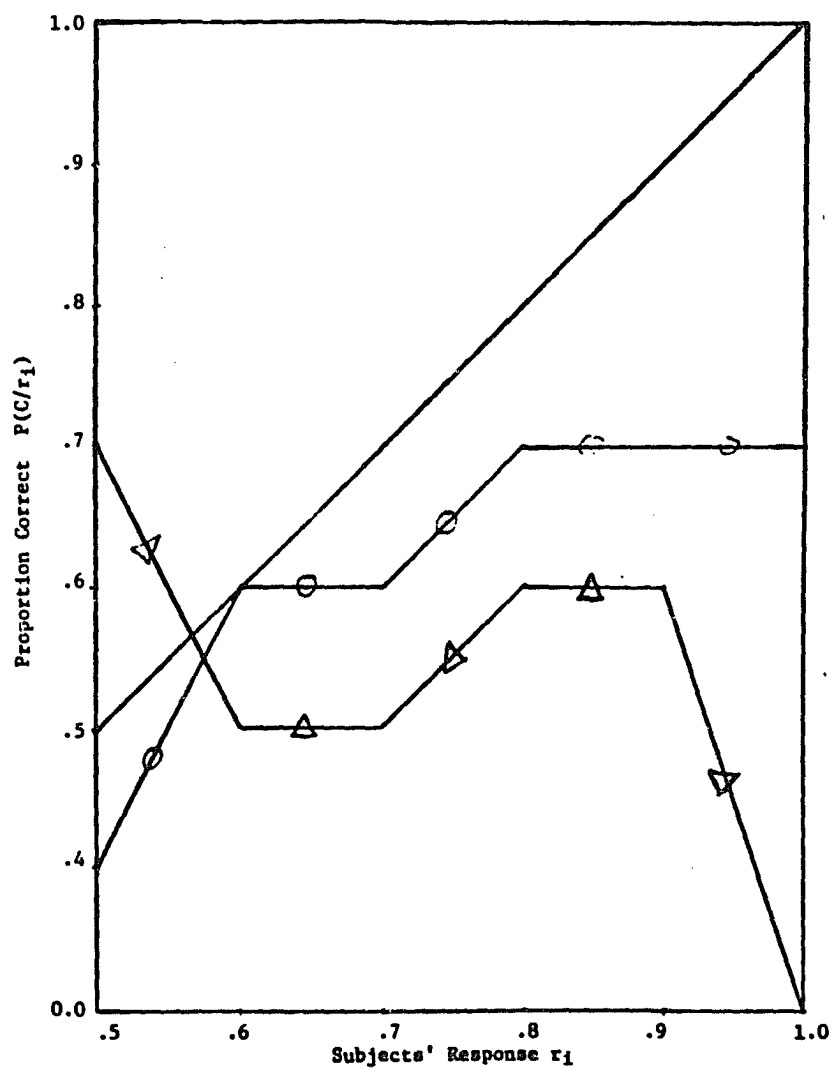


Fig. 7-3. Account Classification Criterion II: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses.

—△— = Good AIC System (ABC)  
 —○— = Bad AIC System (XYZ)

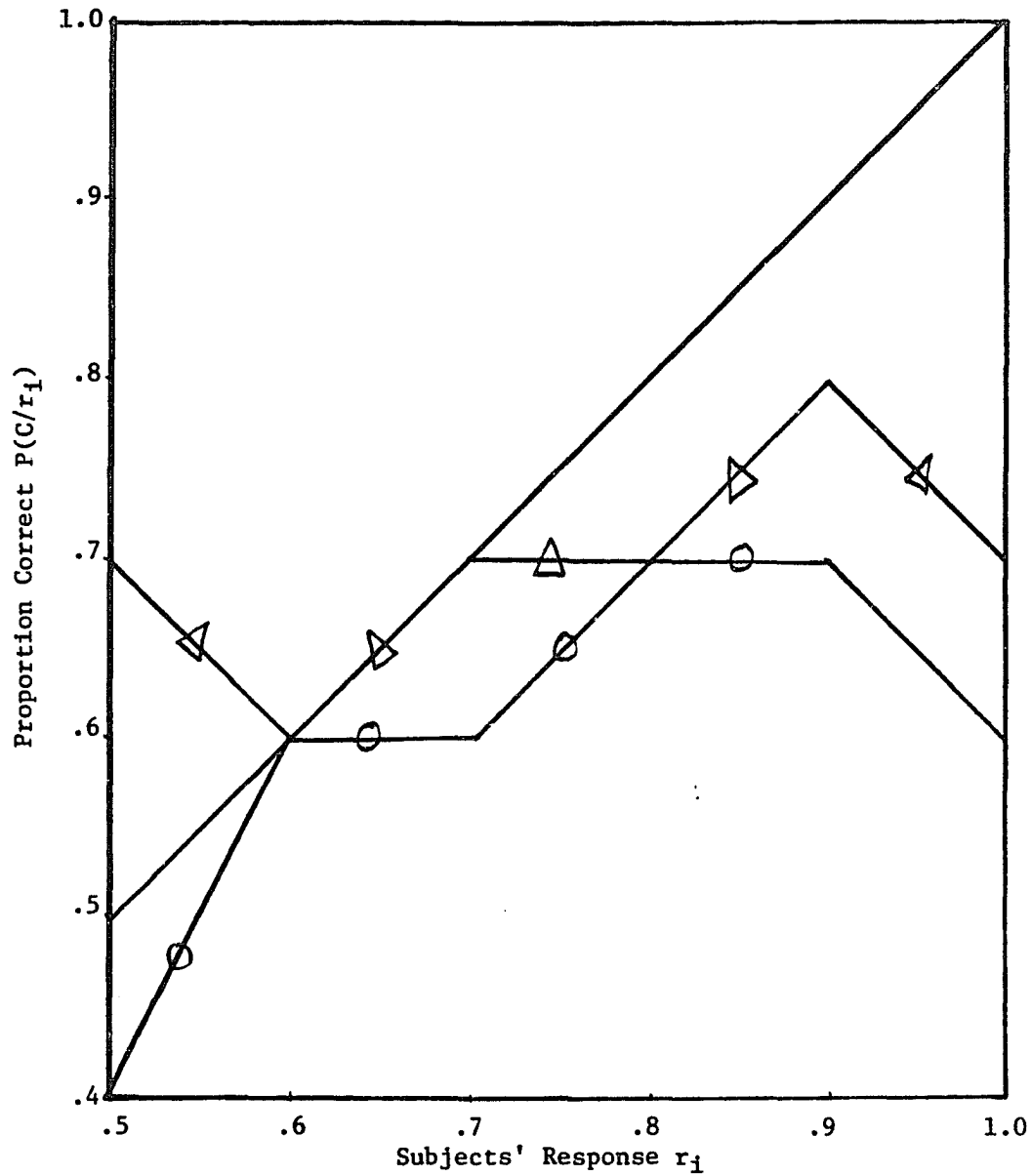


Fig. 7-4. Account Classification Criterion III: Effect of Quality of Accounting Internal Control (AIC) System on Calibration of Subjects' Responses.

$\triangle$  = Good AIC System (ABC)  
 $\circ$  = Bad AIC System (XYZ)

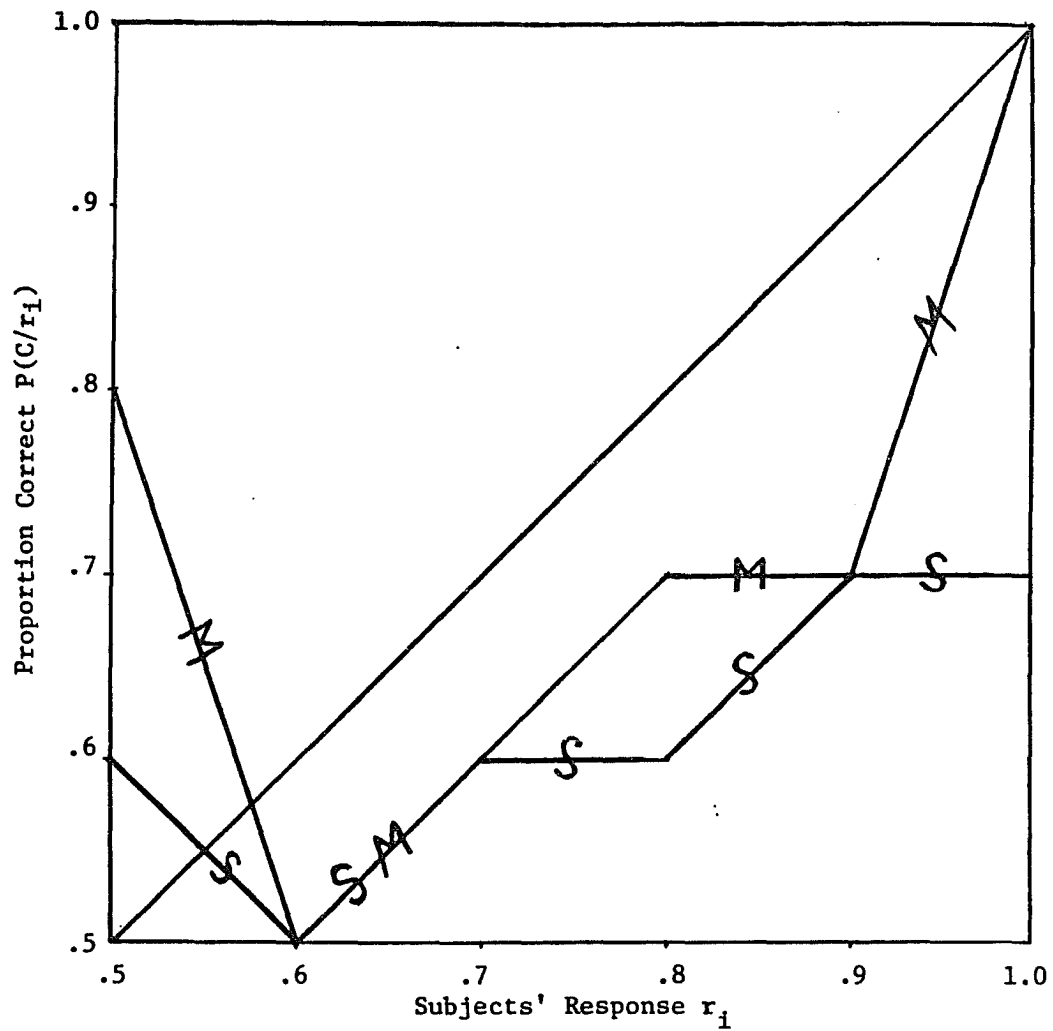


Fig. 7-5. Account Classification Criterion I: Effect of Functional Level on Calibration of Subjects' Response.

—M— = Manager

—S— = Senior

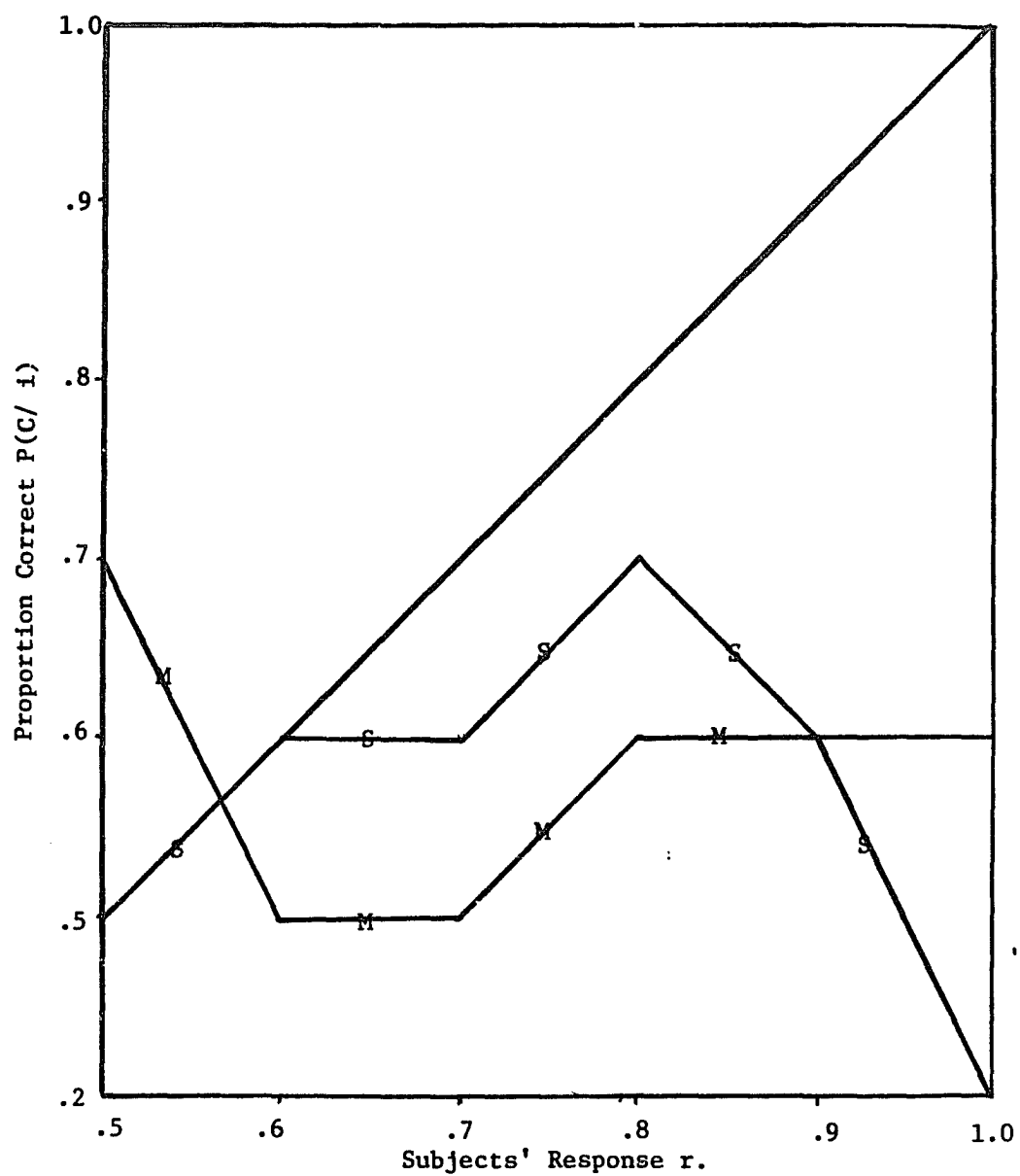


Fig. 7-6. Account Classification Criterion II: Effect of Functional Level on Calibration of Subjects' Responses.

— M — = Manager

— S — = Senior



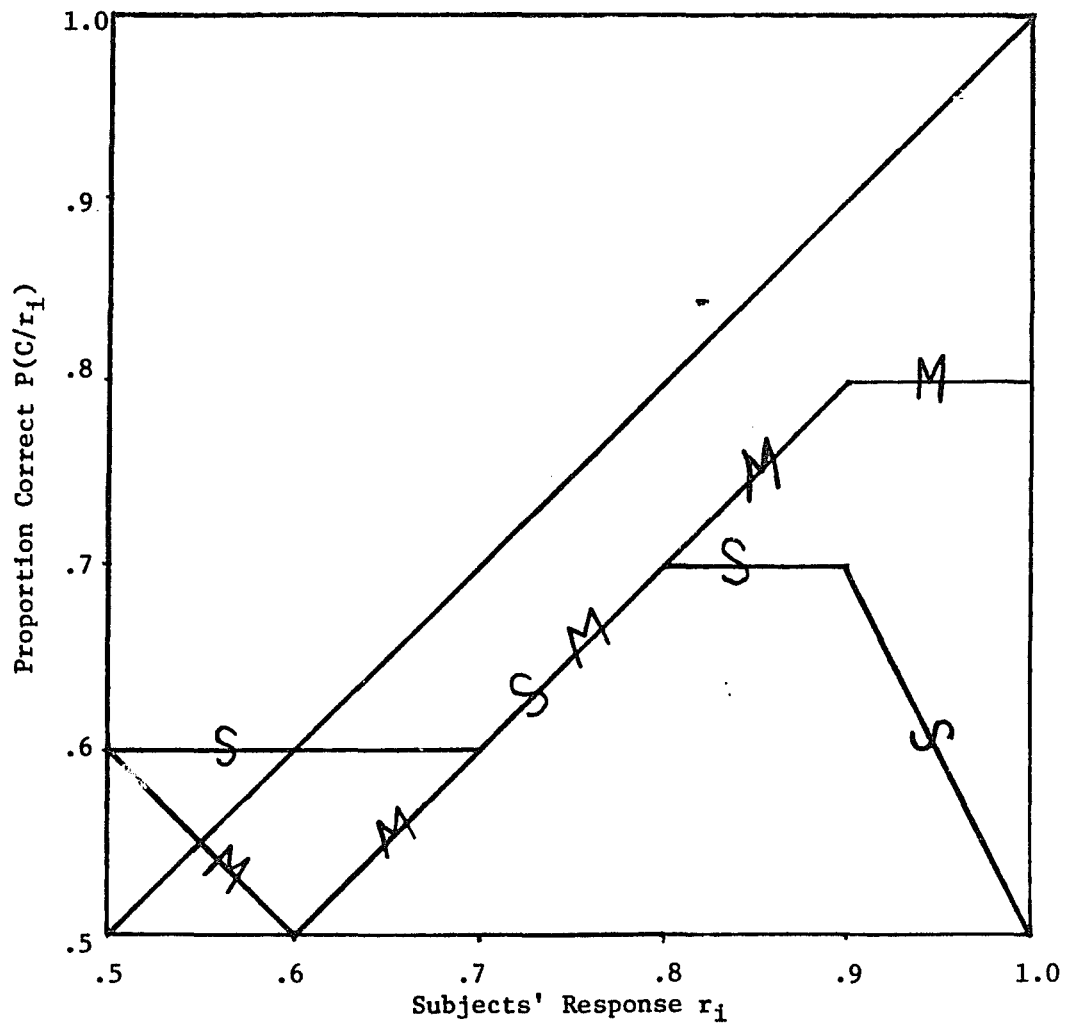


Fig. 7-7. Account Classification Criterion III: Effect of Functional Level on Calibration of Subjects' Responses.

—M— = Manager

—S— = Senior

The following comments seem appropriate regarding the calibration results reported above. It should be noted that the experimental task requires the subjects to state whether an account balance is materially misstated (judgment criterion) and to indicate confidence in the correctness of their responses. Then, in light of these two answers, the subjects were required to indicate whether the account item under consideration will require special audit attention. The subjects' confidence ratings, therefore, apply to the responses under the judgment criterion rather than to those of the action criterion. To evaluate the calibration of auditors' responses under the action criterion, the subjects would have had to provide another set of subjective probability judgments indicating the level of confidence auditors have in the appropriateness of their decisions.

Overall, the findings reported here on the calibration of auditors' responses are consistent with the findings reported in the subjective probability elicitation literature which indicate that subjective probabilities are most often overconfident (see Lichtenstein, et al, 1982). These findings, however, contrast with those reported by Solomon (1982) who reported that the prior probability distributions (PPDs) of the responses of the auditors in his study indicate little tendency towards overconfidence and some tendency

towards underconfidence. Also, this study indicates that the nature of miscalibration of auditors' responses in this study was not sensitive to the state of internal control. It appears, however, that the subjects' responses indicate less overconfidence when the internal control is strong.

The nature of miscalibration of the subjects' responses also provides support for the DVPM. In conformity with the predictions of the model, the base rate  $[P(SN)]$  seems to have no effect on calibration. Similarly, the overconfidence noted with respect to the subjects' responses is consistent with DVPM's prediction, given that the subjects' proportion of correct responses range from 0.57 to 0.73, an indication that the experimental task was difficult.

The relevant literature (e.g., Beck, et al, 1982) suggests that overconfident auditors are likely to collect insufficient audit evidence on which to base their opinion. This study's findings suggest that this relationship may hold only under certain conditions. Given the reported overconfidence of this study's subjects, one expects that they should be less prone to responding "sn" than "n", since in the latter case the auditor will have to perform less tests of details and, hence, obtain less audit evidence. However, the results reported in Table 7-3 indicate that this condition holds

only when the AIC system is adjudged strong. A plausible reason for this finding is the discrepancy between auditors' beliefs and their preferences, which the literature postulates should be identical. However, more research is required before definite conclusions can be drawn.

A plausible cause of the differences in the calibration of auditors' responses noted above and those reported by Solomon (1982) for similar (professional) auditor-subjects might be attributed to differences in the probability elicitation techniques employed in the two studies. In this study, the auditors were provided the book values for each account item, and were required to specify whether the stated book value was (or was not) materially misstated and also to provide the level of confidence in the correctness of their responses. The nature of the task could, therefore, be considered from a signal detection point of view. But in the other studies referred to above, the cumulative distribution function (CDF)-fractile elicitation method was used, in which the subjects were required to state a book value for each fractile category. These differences in the calibration of auditors' responses could, at least in part, be due to differences in the elicitation techniques employed, since other researchers (for example, Chesley 1976) have indicated that the elicitation technique is the greatest

source of variation in the results of studies on subjective probability.

Research Issue Number Seven:

Information Required for PAR Judgments

Table 7-11 presents the (weighted) mean rank (M.R) of information items that the auditor-subjects considered relevant to PAR judgments. The information item considered most relevant for each account item is ranked number 1, with the others ranked in descending order. The table also presents the coefficient of variation (C.V) of these rankings.

The table indicates that there are some information items which the subjects consider relevant for PAR judgments across various account items. These include ratio analysis, history of audit adjustments, quality of internal control system and/or client personnel, and discussions with client management/personnel. Other items of information also considered relevant across many account items include the state of the economy, the nature of the client's industry, and the relationship of the given account item to other related account items. For example, the auditors expect the nature of changes in, say, the Sales Account to be positively related to the nature of changes in the Accounts Receivable account, provided there are no confounding effects of significant

Table 7-11

## Information Items Required for PAR Judgments

Information Item	A C C O U N T		I T E M													
	Sales		Cost of Sales		Income Tax Provision		Inventory		Accounts Receivable		Allowance for Doubtful Accounts		Bad Debt Expense		Accounts Payable	
Information Item	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.	M.R.	C.V.
Ratio Analysis	1	A	2	B	6	F	1	A	2	A					1	A
History of Audit Adjustments	8	G	7	C	4	D	7	G			2	C	3	F	2	B
State of Economy/Industry	3	B	6	E			4	C	5	G	6	C	4	D	5	G
Quality of Internal Control System/Personnel	4	D	4	D	5	F	3	B	4	F	4	F			5	E
Cut-Off Procedures															3	F
Relationship with Related Accounts	6	E	5	F	1	B			3	C			1	A	4	D
Aging of Accounts (or Assets)									1	B	1	A			7	H
Cash Flow Analysis															6	C
Discussions with Management	5	F			8	C	6	E	7	D	5	C	6	C	8	G
Activities in Account/Related Accounts			1	A											6	F
Capitalization/Expensing Policies															1	A
Physical Control Over Assets															2	C
Significance of Foreign Sales/Revenue	2	C			6	F									7	E
Variation from Budget	7	H														
Product Costing System			3	C												
Timing/Permanent Differences					2	A										
Tax Rates (Domestic, Foreign)					3	E										
Inventory Turnover							2	D								
Components of Inventory (EM, WIP, Finished Goods)							5	F								
History of Write-Offs									6	H	3	E	2	B		
Collectibility/Customers' Ability to Pay									8	E	7	B	5	E		

Note: M.R. means Main Rank for each information item per account item.

C.V. means the Coefficient of Variation of the ranking of each piece of information per account item.

changes in other variables like credit policies, the state of the economy, or the customers' ability to pay.

Auditors are generally able to assess the impact of these extraneous variables through discussions with client personnel as well as familiarity with developments in the economy and the client's industry.

This study's findings provide both support and contrast to those of other relevant studies. For example, Kinney (1979) reports that the existence of audit adjustments in an account item in the prior year was the most important indicator of the existence of a material error in the current period's book value. The results reported in Table 7-11 also support the relevance of this information in the context of this study. However, history of audit adjustment was never considered the most important item of information for PAR judgments with respect to any of the account items. It was considered the second most important for only two account items (Allowance for Doubtful Accounts and Accounts Payable), and the least important with respect to three other account items (Sales, Cost of Sales and Inventory). It was not even mentioned as a relevant piece of information for PAR judgments with respect to Accounts Receivable.

Hylas and Ashton's (1982) study provides empirical evidence regarding the perceived diagnostic value of discussions with client management/personnel in the

detection of materially mistated book values. The authors indicate that less rigorous audit procedures such as AR and discussions with client personnel led to the detection of a large proportion of errors. They also indicate that client personnel problems, such as inexperience and insufficient knowledge of accounting, and various types of cut-off or accrual problems are important causes of errors.

This study's findings appear to be consistent with those reported by Hylas and Ashton (1982). For example, discussions with management were required by the subjects for all but one (Cost of Sales) account item. This is not surprising, since the auditor can obtain relevant information regarding the reasonableness of the Cost of Sales book value from other related account items, such as Sales or Inventory. The fact that the subjects ranked activities in related accounts as the most important piece of information required for their PAR judgments with respect to Cost of Sales provides support for this idea. The results also indicate that information regarding the quality of the internal control system and/or personnel was considered necessary for PAR judgments for all account items except Bad Debt Expense. The explanation offered in respect of Cost of Sales regarding the availability of information from other account items also is true with respect to Bad Debt Expense.



The coefficient of variation (C.V.) reported in Table 7-11 was ranked and used as an index of the subjects' degree of consensus on the perceived importance of the information items for PAR judgments with respect to each account item. The lowest (highest) C.V. indicates the highest (lowest) degree of consensus. Of particular interest to this study is auditors' degree of consensus regarding the relative importance to PAR judgments in general the information identified across account items. To provide evidence bearing on this, I used the "Breakdown" procedure (see Nie, et al, 1975) to calculate the descriptive statistics reported in Table 7-12.

The analysis indicate that the mean index of consensus for all information items considered most important for PAR judgments is 1.30 with a standard deviation of 0.483. Similarly, these values are 22.00 and 1.9149 respectively for all information items considered least important. This finding is consistent with the notion that, across account items, the auditor-subjects attained the highest degree of consensus regarding information items considered most important for PAR judgments.

This finding seems encouraging, since it indicates the potential for identifying information items considered most relevant for PAR judgments in general. This augurs well for the development of a behavioral model for

Table 7-12. Auditor's Overall Degree of Consensus Regarding the Relative Importance of Information Items Required for PAR Judgments.

OVERALL INDEX OF CONSENSUS					
Code <sup>+</sup>	Sum	Mean	Standard Deviation	Sum of Squares	Number of Items
1	13.00	1.30	0.483	2.10	10
2	32.00	2.91	1.700	28.91	11
3	36.00	4.00	1.581	20.00	9
4	38.00	3.80	1.549	21.60	10
5	55.00	5.50	1.179	12.50	20
6	56.00	5.09	1.578	24.91	11
7	41.00	5.85	2.268	30.86	7
8	22.00	5.50	1.915	11.00	4

+ The mean rank of information item across all account items.

++ Number of information items within each code level.

auditors' PAR judgments suggested by some researchers noted earlier. However, additional research is required before any generalizable conclusions can be drawn.

## CHAPTER 8

### SUMMARY, LIMITATIONS, IMPLICATIONS OF RESEARCH FINDINGS, AND SUGGESTIONS FOR FURTHER RESEARCH

This final chapter of the dissertation presents (1) the limitations of the study, and (2) a summary of the major findings of the research study, including their likely implications for public accounting practice. The chapter concludes with suggestions for further study.

#### The Limitations of the Study

First, it should be recalled that the auditor-subjects were selected on the basis of availability and willingness to participate. Therefore, in a strict sense, it is inappropriate to generalize the findings of this study beyond the auditors who participated herein.

A second limitation relates to the use of case studies in the experiment. Although an attempt was made to enhance the realism of the experiment as much as possible, some aspects of the real-world AR decision

process were not captured in the experimental setting. For example, the case studies do not capture the real-world reward structures and time pressures in AR tasks. They also did not capture the real-world AR decision process which, as indicated earlier, is multi-stage in nature. Furthermore, the case studies could not present all the information the subjects desired, or what they would normally have had in practice. Nevertheless, it should be noted that the cases were based upon realistic accounting data and information on actual audit clients. In addition, an audit manager in the public accounting firm which provided the data assisted in the determination of the relevant information set for the purpose of the experiment.

A third limitation of the study concerns the absence of an unequivocal criterion for classifying the account items as SN or N. Also, in practice, the audit process determines the "need" to flag an account item for intensive audit. That is, the auditor still may not detect account items which are actually materially misstated. Consequently, there always will be some elements of arbitrariness in classifying account items as SN or N in this type of experimental task.

But, in recognition of the likely effects of the arbitrary classification approach adopted both on the subjects' observed detectability and evaluation of the

external validity of their responses, the subjects' performances were compared under three account classification criteria. The results suggest that, for this study's subjects, the three classification approaches adopted had no significant effect on detectability and judgment biases. However, consistent with the results reported for the simulation experiment, there is a higher degree of variability in the subjects' judgmental accuracy (i.e., Acs) for low P(SN) values.

A final limitation relates to the use of calibration as the only measure of external validity of the subjects' responses. Undoubtedly, this is a narrow perspective from which to evaluate judgments. In addition, some researchers (e.g., Yates, 1982) have suggested that resolution of responses should be preferred to calibration as a measure of external validity of judgments. Nevertheless, calibration seems particularly relevant to evaluating the validity of judgments in the auditing context because of the implications of the nature of miscalibration on audit effectiveness and efficiency. For example, overconfident auditors may collect insufficient sample information on which to base their audit judgments. Similarly, underconfidence implies that auditors might collect more sample information than is required to make audit judgments. Therefore, while admitting that calibration alone is not a complete measure

of validity of judgments, the information from calibration analysis is directly relevant to the purpose of this study.

### The Major Findings of the Study

A summary of the major findings of the research study is presented in this section, followed by a discussion of the implications of the findings.

Overall, the empirical results presented in Chapter 7 indicate that: (1) the detectability of the auditors' responses are relatively high, considering the nature of the data used for the signal detection analysis; (2) the responses were affected by the subjects' biases, suggesting that the auditors confounded their beliefs with their preferences in making their judgments; (3) the auditors' judgments were miscalibrated; and (4) the auditor-subjects displayed a higher degree of consensus regarding the information items considered most relevant to PAR judgments than those considered least important.

The specific research findings by each research issue are summarized as follows:

#### Research Issue One

The Acs reported for the subjects ranged from 0.482 to 0.782. Considering the nature of the data used in the experiment, this finding suggests a reasonably high

degree of detectability by the auditors. Both the state of internal control and functional level appear to have an effect on the subjects' Acs.

#### Research Issue Two

The subjects' responses were affected by their biases. They appear to be concerned more with avoiding, or at least minimizing, the costs associated with incorrect acceptance of materially misstated account balances. Hence, they were more prone to committing efficiency errors by flagging for intensive audit account balances which are fairly presented.

#### Research Issue Three

There is a significant effect of task criterion on the subjects' performances. Their judgmental accuracy was, generally, higher under the judgment criterion, while their decision errors were higher under the action criterion. The predominant type of decision error, also, is the tendency to flag for intensive audit fairly presented account balances. This finding suggests that, when planning for an audit in practice, auditors prefer to play it safe rather than to be sorry. It also suggests the need for an explicit recognition of the effect of implicit loss functions on auditors' judgments under uncertainty.



#### Research Issue Four

The state of the internal control system has an effect on the characteristics of the subjects' responses. In general, the subjects were able to make more accurate judgments when the internal control system was adjudged strong.

#### Research Issue Five

Functional level, in general, appears to have no statistically significant effect on the subjects' responses. However, the seniors committed more effectiveness errors of incorrect acceptance of materially misstated account balances than did the managers. In general, the managers' responses tend to be superior to those of the seniors.

#### Research Issue Six

The subjects' responses were miscalibrated. The subjective probabilities were mostly overconfident for all account classification criteria.

#### Research Issue Seven

Consistent with the findings reported in earlier studies, simple AR procedures such as ratio analysis, scanning, and comparisons amongst data, were the ones most preferred by the auditor-subjects to facilitate their

preliminary analytical review judgments. Furthermore, the auditors displayed a higher degree of consensus regarding the information items considered most important for facilitating PAR judgments than those considered least important.

Implications of Research  
Findings and Suggestions  
For Further Research

This section discusses the implications of the research findings noted above, upon which suggestions for further research are made.

Despite the pervasive role of auditor judgment in PAR procedures, existing research has focused only on evaluating the performance of alternative statistical AR models. This research study has demonstrated that auditors are reasonably good at identifying, at the onset of an audit, account items which are likely to be materially misstated. However, more research of this nature is called for before definite conclusions can be drawn.

Also required is more research comparing the ability of auditors and the ability of statistical models at identifying, at the onset of an audit, account items which are materially misstated. But, as stated earlier, statistical models merely supplement human judgments in practice. Of interest, therefore, should be evidence

regarding the incremental value of statistical models at enhancing the accuracy of auditors' PAR judgments.

This study's findings indicate that the behavior of subjects in an experimental task could be affected by the perceived consequences of their judgment errors. For example, auditors' responses under the judgment criterion significantly differ from their responses under the action criterion. Under the former criterion, the subjects were required merely to state whether an account balance was or was not materially misstated, without regard to the consequences of their responses. Under the latter criterion, the subjects were required to indicate what specific actions they would have taken in practice. In this case, one expects the subjects to be conscious of and sensitive to the consequences of their decisions.

Earlier studies have confounded these two aspects of judgments in experimental tasks. Yet, this study's findings suggest that the subjects' decision rules might differ significantly for each of these situations. For example, they suggest that the subjects' decision rules are more stringent under the action criterion than under the judgment criterion. The nature of the difference in stringency of decision rules also suggests that auditors are more likely to be risk-averse when actual audit decisions are being made.

It appears, therefore, that auditors' beliefs may not be independent of their preferences. This observation has an implication for the application of the Bayesian approach to audit decision making, which postulates that auditors' beliefs be independent of their preferences. However, further research is required before any definite conclusions can be drawn.

A useful byproduct of the distinction between judgments and decisions is the finding that, although auditors might perceive an account balance to be not materially misstated, they may still decide to slate such an account for intensive audit. This finding provides a useful insight into the auditors' decision-making strategies, and paves the way for a better understanding of the behavioral factors which affect auditors' judgments. For example, this revelation helps in explaining why auditors might be more prone to committing the error of slating for intensive audit fairly presented account balances.

In a wider context, this finding has implications for inferences made in the human information processing (HIP) literature which merely indicate that judges perform less optimally than normative models. A better understanding of the specific decision strategies employed by the judges in each experimental task should help to explain the reasons for the alleged suboptimal behavior,

assuming that the given normative models with which judges' performances were being compared were appropriate.

The identical nature of (mis)calibration of the subjects' responses under the three account classification criteria provides additional support for the DVPM developed by Ferrell and McGoey (1980). This result also is consistent with the findings reported by Smith and Ferrell (1981), who applied the DVPM to a task in which the subjects provided responses to a set of general knowledge (almanac) items. Also, as indicated earlier, the model predicts that, in an experimental task in which subjects were to decide whether a proposition is true or false and subsequently to give the subjective probability that the decision was correct, base rate should have no effect on calibration. It also elucidates the effect of task difficulty on calibration, as predicted by DVPM.

This model thus suggests a means of removing some extraneous factors which may affect the evaluation of the nature of (mis)calibration of subjective judgments. Hence, the feasibility of its adoption for evaluating auditor judgments under uncertain conditions should be investigated.

The information items that auditors consider relevant for PAR judgments are consistent with those of other studies. In addition, there was a low degree of variability in auditors' rankings of information items

Considered most relevant for PAR judgments. This finding seems promising for the identification of information items which may be useful for developing a behavioral model of auditors' PAR judgments. However, more research is required before any generalizable conclusions can be drawn.

In view of the above, the following recommendations are made for further research: (1) more research using the signal detection model in the AR context is suggested as a basis for evaluating the usefulness of the model for analyzing accounting and auditing judgments; (2) research should be conducted on the effect of the reward structure on auditors' attitudes towards risk and judgment biases; and (3) further research should be conducted on the relative importance of information considered necessary for PAR tasks, possibly by type of firm and/or industry, to enhance the development of a behavioral model for AR judgments suggested by some researchers.

## APPENDIX

## APPENDIX A

### EXPERIMENTAL MATERIALS

- 1) Letter Informing Participating CPA Firms About the Date of Administration of the Experiment
- 2) Introduction
- 3) Background Information on the Electronics Industry
- 4) ABC, Inc.: Background Information
- 5) ABC, Inc.: Response Sheets
- 6) XYZ, Inc.: Background Information
- 7) XYZ, Inc.: Response Sheets
- 8) Subjects' Background Data





**THE UNIVERSITY OF ARIZONA**  
TUCSON, ARIZONA 85721

COLLEGE OF BUSINESS AND  
PUBLIC ADMINISTRATION  
DEPARTMENT OF ACCOUNTING

July \_\_, 1982

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
Dear Mr./Ms. \_\_\_\_\_:

This letter is a follow-up to the telephone discussion you recently had with Dr. Ira Solomon on your firm's participation in my audit research study. I am a doctoral student in Accounting at The University of Arizona, and Dr. Solomon is supervising this research project.

As already discussed with Dr. Solomon, the experiment is to be administered on Thursday, July \_\_, 1982 at \_\_\_\_ A.M./P.M., and will proceed as follows. Initially, I will orientate the participants through a brief discussion of the focus and objectives of the study. I also will provide limited training by going through an example of the experimental task identical to that contained in the actual experimental materials. Having answered the questions posed by the participants to clarify any problems arising from the orientation, I will then provide them with the actual experimental materials.

I suggest that each participant bring a calculator for minor computations and materials for making personal notes. I also request that a room be made available which is large enough to accommodate all the participants at once.

Your assistance in these matters and your willingness to provide participants from your office are greatly appreciated. I look forward to meeting you shortly.

Sincerely,

Ademola Ariyo  
(Doctoral Student)

AUDITING RESEARCH STUDY:

AUDITOR-JUDGMENTS IN  
PRELIMINARY ANALYTICAL  
REVIEW

### INTRODUCTION

This research project is concerned with how professional auditors' preliminary analytical review judgments, based on limited information available at the onset of an audit, are used to identify account items which are likely to be materially misstated, and, hence, may require special audit attention. As you read over the following pages and provide the required responses, you are to assume the role of the audit team member who is responsible for allocating audit efforts on the engagement. You may be assured that neither you nor your firm will be identified as a participant in the study. Thank you for your participation.

Two cases have been developed from data provided by a national public accounting firm on two different audit clients in the electronics industry. On the pages that follow, you will be presented with (a) general operating and financial information on the two independent audit clients, (b) background information on the electronics industry, and (c) a set of unaudited 1981 account book values for each firm. Given this limited information, you will be asked to respond to each of the following questions for each account book value:

- 1) Do you think this account book value is materially misstated?

Yes

No

- 2) What is your subjective probability regarding the correctness of your answer to (1)?

.5                  .6                  .7                  .8                  .9                  1.0

---

(I am completely uncertain whether my answer is correct or incorrect.)

(I am absolutely certain that my answer is correct.)

- 3) Do you think this account book value will require special audit attention?

Yes

No

- 4) Please list below in descending order of importance five (5) relevant items of information you normally require in practice for making the type of judgment in question (1) for this account item.

These questions are reproduced later at the appropriate points.

You are free to go back and change any of your responses anytime you feel like doing so. The questionnaire ends with a set of debriefing questions.

BACKGROUND INFORMATION ON  
THE ELECTRONICS INDUSTRY

The electronics industry has witnessed a sustained dramatic growth in recent years, with the demand for all types of electronic products rising steadily for the past 10 years. Domestically, the government, especially the Department of Defense, usually accounts for about 50% of all shipments of electronics products. The projected increases in defense expenditures by the public sector, therefore, augur well for the future growth of the industry.

Traditionally, the industry has been resilient against recessionary trends. The industry's growth in 1981 has, however, been adversely affected by the continued economic downturn, such that its 1981 growth in real terms was about one half of one percent (0.5%). Nevertheless, the anticipated economic upturn, coupled with the ever broadening application of electronic products in various sectors both domestically and internationally, suggest a bright future for the industry as a whole.

The electronics industry is both highly competitive and characterized by rapid technological changes such that, as a consequence, there is a high rate of inventory obsolescence. Intensified domestic and international competition also has resulted in a drastic reduction in the gross margins for most electronic products. The ability of many firms to maintain their market shares in the future may, therefore, depend significantly on product quality and lower prices.

## ABC, Inc.

## BACKGROUND INFORMATION

## General

ABC, Inc. is a privately-held 24-year old company, whose main activity is the application of electronic technology to industrial, commercial, and military markets worldwide. All manufacturing activities take place at the company's facilities in the U.S. The company's annual sales rose from about \$16,000,000 in 1976 to about \$61,000,000 in 1981, and the company has exceeded its targeted annual growth rate of 25% in earnings for almost ten years. ABC's growth rate both in sales and in earnings compare favorably to those of its competitors.

As of 1981, half of the total annual sales is accounted for by foreign subsidiaries, nearly all of which are in Europe. The company has over 10,000 customers representing a healthy balance of markets, geographics, industries, and end products worldwide. The largest customer accounts for only about 2% of the firm's total annual sales. Both the level of operations and cash flows of some of these customers are, however, being adversely affected by the current downturn in the economy of the U.S. and those of the foreign customers' countries.

Because of the current worldwide economic downturn, especially in those countries in which it operates, ABC's rate of growth of sales backlog declined in 1981, resulting in a lower inventory turnover. Some workers also were laid off. There was, however, only a slight decline in the growth rate of the company's sales (relative to that of

prior years) because the full impact of the reduction in labor force was partially offset by an overall increase in worker productivity. This achievement was attributed to the effectiveness of the company's training center and education program aimed at increased productivity and employee development.

#### Internal Control

ABC's internal control system is, generally, very good, with competent accounting and EDP personnel, as well as adequate segregation of duties. The company typically acts promptly on the external auditor's recommendations contained in the management's letter. Consequently, significant improvements have been made in accounting problem areas identified earlier by the auditors. ABC also seeks advice from the auditors before embarking on any program that is of accounting significance. For example, the company requested internal control advice before implementing a new cash management system to accelerate collection of accounts receivables. The company also recently initiated a material resource planning model as part of its continuing efforts to enhance the management effectiveness.

Historically, the internal control subsystems have been found to be reliable, with compliance exception rates ranging from 1% to 3%. However, ABC has neither an internal audit function, nor an audit committee.

The following pages present financial and operating data, supplemental notes to the financial data, and response sheets for you to indicate your analytical review judgments.

## ABC, INC.

FINANCIAL AND OPERATING DATA

<u>ACCOUNT ITEM</u>	<u>AUDITED</u>		<u>UNAUDITED</u>
	<u>1979</u>	<u>1980</u>	<u>1981</u>
	(Thousands of \$)		(Thousands of \$)
SALES: Domestic	22,504	28,620	30,684
International	20,773	27,496	30,684
	<u>43,277</u>	<u>56,116</u>	<u>61,368</u>
COST OF SALES	31,838	40,150	42,830
INCOME BEFORE TAXES	3,511	3,709	4,065
INCOME TAX PROVISION	1,701*	1,070	2,236 <sup>(a)</sup>
NET INCOME	1,810	2,639	1,829
INVENTORY:			
Finished Goods	2,384	3,325	4,378
Work-in-Process	2,431	4,287	4,881
Raw Materials	<u>3,218</u>	<u>4,614</u>	<u>4,028</u>
	8,022	12,226	13,287
Excess Inventory Reserve <sup>(b)</sup>	<u>(177)</u>	<u>-0-</u>	<u>(1,350)</u>
NET INVENTORY	<u>7,856</u>	<u>12,226</u>	<u>11,937</u>
Inventory Expenses as Obsolete	-0-	152	307
ACCOUNTS RECEIVABLE:			
Consolidated Balance	8,558	10,145	10,702
Allowance for Doubtful Accounts	<u>(60)</u>	<u>(60)</u>	<u>(60)</u>
NET ACCOUNTS RECEIVABLE	<u>8,498</u>	<u>10,085</u>	<u>10,642</u>
BAD DEBT EXPENSE <sup>(c)</sup>	25	27	52
ACCOUNTS PAYABLE	3,939	4,125	4,110
PLANT, PROPERTY & EQUIPMENT (PPE)			
Land	191	191	191
Total Depreciable Assets	<u>14,218</u>	<u>20,040</u>	<u>20,466</u>
	14,409	20,231	20,657
Accumulated Depreciation	<u>(3,476)</u>	<u>(5,154)</u>	<u>(6,666)</u>
NET PPE	<u>10,933</u>	<u>15,077</u>	<u>13,991</u>
DEPRECIATION EXPENSE	1,810	2,808	2,606



SUPPLEMENTAL NOTES TO FINANCIAL DATA

- a) The domestic tax rate was 46%, but the tax rate in one of ABC's most important foreign subsidiaries was about 58% for 1981.
- b) "Excess Inventory Reserve" represents inventory reserve for slow-moving and obsolete inventories of finished goods and raw materials which exceed 12-month projected usage for related products.
- c) The Accounts Receivable Aging Analysis is as follows:

	<u>A U D I T E D</u>				<u>UNAUDITED</u>	
	<u>1979</u>		<u>1980</u>		<u>1981</u>	
	\$(000s)	%	\$(000s)	%	\$(000s)	%
Current	4,536	53.0	6,493	64.0	5,672	53.0
31-60 days	2,738	32.0	2,029	20.0	2,676	25.0
61-90 days	599	7.0	812	8.0	856	8.0
Over 90 days	<u>685</u>	<u>8.0</u>	<u>811</u>	<u>8.0</u>	<u>1,498</u>	<u>14.0</u>
TOTAL	<u>8,558</u>	<u>100.0</u>	<u>10,145</u>	<u>100.0</u>	<u>10,702</u>	<u>100.0</u>

ABC, INC.: RESPONSE SHEETS

Please respond to the questions provided  
on the following pages for each of the  
following ten (10) account book values  
of ABC, Inc.

Thank you.

ABC, INC.

ACCOUNT ITEM: SALES (1981)

The unaudited Sales book value for 1981 is \$61,368,000.

1) Do you think this account book value is materially misstated?

Yes

No

2) What is your subjective probability regarding the correctness of your answer to (1)?

<u>.5</u>	<u>.6</u>	<u>.7</u>	<u>.8</u>	<u>.9</u>	<u>1.0</u>
-----------	-----------	-----------	-----------	-----------	------------

(I am completely uncertain whether my answer is correct or incorrect.)

(I am absolutely certain that my answer is correct.)

3) Do you think this account book value will require special audit attention?

Yes

No

4) Please list below in descending order of importance five (5) relevant items of information you normally require in practice for making the types of judgment in question (1) for this account item.

1.

2.

3.

4.

5.

## XYZ, INC.

## BACKGROUND INFORMATION

## General

XYZ, Inc., a new audit client, is a closely-held publicly-traded company, in which members of the same family own about 44% of the voting shares. Since its incorporation about thirty years ago, XYZ has engaged in the design, development, manufacture, and sale of computer related products. Its annual sales have risen from about \$10 million in 1976 to about \$19 million in 1981. Foreign subsidiaries accounted for about 15% of total sales in 1981, while one customer alone accounted for about 21% of domestic sales for the same fiscal year. All manufacturing activities, however, take place at the company's facilities in the U.S.

Both the adverse effects of continued inflation and reduced margins on some of the company's products, caused by intensified domestic and international competition, have accounted for the downward trend in profitability for the last few years. Also, the rate of growth of the firm's sales is lower than those of its competitors. Furthermore, because of increases in short term borrowings, interest expenses have increased substantially for the past two years, increasing by about 27% from fiscal 1979 to fiscal 1980, and by about 30% from fiscal 1980 to fiscal 1981. The company, however, has been able to achieve a gradually declining rate of growth in expenses relative to that of sales, through a combination of cost reducing measures and efforts aimed at improving employee productivity.

Most of XYZ's competitors are larger and have greater financial resources. XYZ believes that its ability to continue to compete successfully in its present markets is dependent on a combination of product quality, service, and price. Management also believes that the company has sufficient financial resources and personnel to maintain its competitive position in its present business.

#### Internal Control

XYZ's internal control is on the lower end of a spectrum of internal controls systems upon which the external auditor may place reliance. In 1981, there was a large turnover of accounting personnel, including both the controller and the accounting supervisor. Their respective successors are less familiar with the electronics industry, and the new accounting supervisor is yet to fully comprehend XYZ's accounting practices.

Discussions with the predecessor auditors indicate that historically, XYZ booked several adjustments as a result of the external audit. Most such adjustments could be attributed to unintentional errors (mostly relating to purchases and costing of inventories), or nonconformity with company accounting principles on a consistent basis (especially relating to capitalizing versus expensing certain expenditures). The firm also rarely seeks the advice or services of the auditors on issues of accounting significance. For example, XYZ usually files quarterly reports without prior review. Hence, such reviews typically are performed retrospectively.

The EDP systems are out of date, and new ones are urgently required. In addition, XYZ's cost accounting system is unreliable. Although there appears to be a reasonable segregation of duties between related employees, four members of the same family hold key corporate positions enhancing the likelihood of management override of the internal control system. The firm has no internal audit function, but does have an active three-member audit committee (two of which are nonemployee directors).

The following pages present the financial and operating data, supplementary notes to financial data, and response sheets relating to each account book value of XYZ, Inc.

## XYZ, INC.

## FINANCIAL AND OPERATING DATA

<u>ACCOUNT ITEM</u>	<u>AUDITED</u>		<u>UNAUDITED</u>
	<u>1979</u>	<u>1980</u>	<u>1981</u>
	(Thousands of \$)	(Thousands of \$)	(Thousands of \$)
SALES: Domestic	11,918	10,995	16,168
International	<u>4,286</u>	<u>4,646</u>	<u>2,844</u>
	<u>16,204</u>	<u>15,641</u>	<u>19,012</u>
COST OF SALES	<u>10,967</u>	<u>11,851</u>	<u>14,338</u>
INCOME BEFORE TAXES	1,429	255	289
INCOME TAX PROVISION	597	47	110
NET INCOME	832	208	179
INVENTORY:			
Finished Goods	283	204	481
Work-in-Process	555	874	912
Raw Materials	<u>2,624</u>	<u>3,158</u>	<u>2,647</u>
	<u>3,462</u>	<u>4,236</u>	<u>4,040</u>
Inventory Valuation Reserve <sup>(a)</sup>	<u>-0-</u>	<u>(185)</u>	<u>-0-</u>
NET INVENTORY	<u>3,462</u>	<u>4,051</u>	<u>4,040</u>
Inventory Expensed as Obsolete	15	189	360
ACCOUNTS RECEIVABLE:			
Consolidated Balance <sup>(b)</sup>	3,344	3,045	3,775
Allowance for Doubtful Accounts	<u>(77)</u>	<u>(77)</u>	<u>(115)</u>
NET ACCOUNTS RECEIVABLE	<u>3,267</u>	<u>2,968</u>	<u>3,660</u>
BAD DEBT EXPENSE	26	1	61
ACCOUNTS PAYABLE	1,354	1,407	1,080
PLANT, PROPERTY & EQUIPMENT (PPE)			
Land	302	302	302
Depreciable Assets	<u>5,040</u>	<u>5,156</u>	<u>5,286</u>
	<u>5,342</u>	<u>5,458</u>	<u>5,588</u>
Accumulated Depreciation	<u>(1,490)</u>	<u>(1,845)</u>	<u>(2,186)</u>
NET PPE	<u>3,852</u>	<u>3,613</u>	<u>3,402</u>
DEPRECIATION EXPENSE	311	371	502

SUPPLEMENTAL NOTES TO FINANCIAL DATA

- a) "Inventory Valuation Reserve" represents reserve for slow-moving inventories of finished goods and raw materials which would have to be sold below current prices because of obsolescence.
- b) The Accounts Receivable Aging Analysis is as follows:

	<u>A U D I T E D</u>				<u>UNAUDITED</u>	
	<u>1979</u>		<u>1980</u>		<u>1981</u>	
	\$(000s)	%	\$(000s)	%	\$(000s)	%
Current	2,508	75.0	1,857	61.0	2,227	59.0
31-60 days	602	18.0	731	24.0	679	18.0
61-90 days	134	4.0	244	8.0	189	5.0
Over 90 days	<u>100</u>	<u>3.0</u>	<u>213</u>	<u>7.0</u>	<u>680</u> *	<u>18.0</u>
	<u>3,344</u>	<u>100.0</u>	<u>3,045</u>	<u>100.0</u>	<u>3,775</u>	<u>100.0</u>

\* Includes \$363,000 of long-term extended Accounts Receivables. Otherwise, the aging percentage for this category would have been 8%.



XYZ, INC.: RESPONSE SHEETS

Please respond to the questions provided  
on the following pages for each of the  
following ten (10) account book values  
of XYZ, Inc.

Thank you.

XYZ, INC.

ACCOUNT ITEM: SALES (1981)

The unaudited Sales book value for 1981 is \$19,012,000.

- 1) Do you think this account book value is materially misstated?

Yes

No

- 2) What is your subjective probability regarding the correctness of your answer to (1)?

.5	.6	.7	.8	.9	1.0
----	----	----	----	----	-----

(I am completely  
uncertain whether  
my answer is cor-  
rect or incorrect.)

(I am absolutely  
certain that my  
answer is  
correct.)

- 3) Do you think this account book value will require special audit attention?

Yes

No

- 4) Please list below
- in descending order of importance
- five (5) relevant items of
- information you normally require in practice
- for making the type of judgment in question (1) for this account item.

1.

2.

3.

4.

5.

## BACKGROUND DATA

I. GENERAL

Present Position in firm:

Are you a CPA?                      Yes              No

II. EXPERIENCE

How long have you been working as an independent auditor?

\_\_\_\_\_ months

How many different audit engagements (count each year's examination of a given client as a separate engagement) have you worked on:

Number of Audit Engagements

	<u>Electronics</u>	<u>Other</u>
<u>Total</u>	<u>Industry</u>	<u>Industries</u>

On how many of these engagements have you been directly involved in preliminary analytical review judgements:

	<u>Electronics</u>	<u>Other</u>
<u>Total</u>	<u>Industry</u>	<u>Industries</u>

How would you describe the degree to which computer statistical packages (e.g., regression analysis) are employed in analytical review in your CPA firm? (Circle one)

Low	Medium	High
-----	--------	------

On what percentage of your analytical review tasks have you employed these computer statistical packages?

\_\_\_\_\_ %

### III. TRAINING

College Education (circle the highest level attained):

Bachelors

Masters

Doctoral

How many college courses have you taken in auditing? \_\_\_\_\_

How many college courses have you taken in probability  
and statistics? \_\_\_\_\_

Please describe briefly any other relevant training  
(relating to analytical review) you have received:

### IV. MATERIALITY THRESHOLD

(a) Did you employ any particular materiality threshold  
for your judgments?

Yes

No

(b) If your answer to (a) is "yes", did you use a common  
materiality threshold for all account items?

Case

Yes

No

ABC

XYZ

(c) If you answer "yes" to (b), please state the threshold  
value you used:

ABC \_\_\_\_\_

XYZ \_\_\_\_\_

If your answer to (b) is "no", please specify the materiality  
threshold used for your judgments relating to each account  
item:

<u>Account Item</u>	<u>Materiality Threshold</u>	
	<u>ABC</u>	<u>XYZ</u>
Sales		
Cost of Sales		
Income Tax Provision		
Inventory		
Accounts Receivable		
Allowance for Doubtful Accounts		
Bad Debt Expense		
Accounts Payable		
Fixed Assets		
Depreciation Expense		

## APPENDIX B

An overview of Grey & Morgan's model and  
the Procedures for Analyzing the Study's  
Data.

# AN OVERVIEW OF GREY AND MORGAN'S STATISTICAL MODEL

The computer program used to analyze this study's data estimates the maximum likelihood estimates (MLE) of the parameters of a normal distribution, and is written by Grey and Morgan (1972). I present an overview of the model as follows.

The model assumes that the stimulus is normally distributed. It also assumes that the distribution of the noise stimulus (N) is of the form  $F_n(x) = F(x)$ , and that of signal-plus-noise (SN) is of the form  $F_{sn}(x) = F(Ex-A)$ , where F is normal.

The ratio of the standard deviation of SN to N is then  $1/B$ ; that is,

$$\frac{1}{B} = \frac{\text{Standard deviation (SN)}}{\text{Standard Deviation (n)}}$$

The separation of the mean (which is equivalent to detectability) of the distributions for SN and N can then be defined by  $A/B$ . That is,

$$U_{sn} - U_n = A/B$$

Hence, if one assumes that the distribution of N is standard normal, that is,  $F_n(x) = F(x)$  is normally

distributed with mean zero and standard deviation of 1, then the standard normal distribution of SN, that is  $F_{sn}(x) = F(Bx-A)$ , is normally distributed with mean of  $A/B$  and standard deviation of  $(1/B)$ .

#### Procedures for Data Analysis

As indicated in Chapter 5, the rating scale experimental approach was used to elicit responses from each of the auditor-subjects, who specified a level of confidence in the correctness of their responses (i.e., answers) on a half-range  $[-.5, 1]$  probability scale.

Since the subjects responded to both types (N and SN) stimuli on the half-range probability scale, their responses were coded in a way consistent with the underlying distributions assumed for each type of stimulus. This leads to eleven response categories shown in Figure AE-1.

The hatched areas represent the data points, in which

A represents the false alarms;  
 B represents the correct rejections;  
 C represents the hits, and  
 D represents the misses.

The probability levels were converted to response categories, while data points were re-grouped into N and SN categories. The format for the data set derived for



		1.0	.9	.8	.7	.6	.5	.6	.7	.8	.9	1.0
N	S											
	N											
SN	S											
	N											

A = False Alarms  
B = Correct Rejects

C = Hits  
D = Misses

Fig. AB-1. The Symmetric Probability Scale Used to Code the Subject's Responses.

	1	2	3	4	5	6	7	8	9	10	11
N											
SN											

Fig. AB-2. Response Categories for Coding the Subjects' Responses\*.

\* Categories (1) and (2) were collapsed for the purposes of data analysis.

each level of data analysis [i.e., overall, by (a) quality of AIC, and (b) functional level] is shown in Figure AB-2.

Since the computer program used allows for a minimum of three, and a maximum of ten, response categories, categories (1) and (2) were collapsed, leaving a total of ten used for the data analysis.

For each response category, which is equivalent to a cut-off (say,  $Z_m$ ), the corresponding value of the likelihood ratio [i.e.,  $Beta(Z_m)$ ] was computed as follows:

$$Beta(Z_m) = F_{sn}(Z_m)/F_n(Z_m)$$

for  $m = 1, 2, \dots, 10$ .

To derive the area under the ROC curve (i.e.,  $A_c$ ), which is equivalent to a nonparametric measure of detectability, Simpson's Rule was used to integrate the area covered by the curve which passes through the locus of points representing the hit and the false alarm rates derived for each  $Z_m$ .

To compute the median Beta [say,  $Beta(Z^*_m)$ ] which represents the subjects' cut-off criterion for separating the N and the SN stimuli, I used the average of the Beta values of the cut-off points  $Z_4$  and  $Z_5$ . That is,

$$Beta(Z^*_m) = Beta[(Z_4 + Z_5)/2].$$

This value represents the subjects' index of bias from which the  $K$  values reported in the study were computed.

## BIBLIOGRAPHY

- Abdel-Khalik, R.A. and K.M. El-Shesai, "Information Choice and Utilization in an Experiment on Default Prediction", Journal of Accounting Research, Autumn 1980: 325-42.
- Adams, J.K. and P.A. Adams, "Realism of Confidence Judgments," Psychological Review, 58, 1961: 33-45.
- Akresh, A.D. and W.A. Wallace, "The Application of Regression Analysis for Limited Review and Audit Planning," University of Illinois Audif Conference, 1980.
- American Institute of Certified Public Accountants, "Exposure Draft on Proposed Statement on Auditing Standards: Analytical Review Procedures," May 15, 1978.
- Angus, R.C. and T.C. Daniel, "Applying Theory of Signal Detection in Marketing: Product Development and Evaluation," American Journal of Agricultural Economics, 56, 1974: 573-577.
- Ashton, R.H., "An Experimental Study of Internal Control Judgments," Journal of Accounting Research, Spring 1974: 143-57.
- Banks, W.P., "Signal Detection Theory and Human Memory," Psychological Review, 74, 1970: 81-99.
- Barr-Brown, M. and M.J. White, "Sex Differences in Recognition Memory", Psychonomic Science, 25, 1971: 75-76
- Beck, P.J., I. Solomon, and L.A. Tomassini, "Subjective Prior Probability Distributions and Audit Risk", Unpublished Working Paper, 1982.
- Biggs, S.C. "Perspectives in Auditing for the 1980s and Beyond," Accounting Research Convocation, University of Alabama, 1981.
- Birnbaum, M.H., "The Devil Rides Again: Correlation as an Index of Fit," Psychological Bulletin, 79, 1973: 239-242.
- Blocher, E. and R.P. Moffle, "A Signal Detection Model for Analyzing Accounting and Auditing Judgments," Unpublished, University of North Carolina, 1982.
- Blocher, E., R.S. Esposito, and J.J. Willingham, "A Study of Auditor Judgments Concerning the Nature and Extent of Analytical Review in Auditing Payroll," Unpublished, University of North Carolina, 1981.

- Brier, G.W., "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, 75, 1950: 1-3.
- Brown, C., "Human Information Processing for Decisions to Investigate Cost Variances," Journal of Accounting Research, 19, 1981: 62-84.
- Brown, T.A. and H. Shuford, Jr., "Quantifying Uncertainty Into Numerical Probabilities for the Reporting of Intelligence," The Rand Corporation, R-1185-ARPA, 1973.
- Bush, R.R., "Estimation and Evaluation," in Bush, R.R., E. Galanter, and R.D. Luce (eds.), Handbook of Mathematical Psychology, Vol. 1, (New York: Wiley, 1973).
- Chapman, C.R. and B.W. Feather, "Sensitivity to Phobic Imagery: A Sensory Decision Theory Analysis," Behavior Research and Therapy, 9, 1971: 161-168.
- Charness, N., "Human Chess Skill," in P. W. Frey (ed.), Chess Skill in Man and Machine, (New York: Springer-Verlag, 1976): 34-53.
- Chase, W.G. and H. A. Simon, "The Mind's Eye in Chess," in W. G. Chase (ed.), Visual Information Processing, (New York: Academic Press, 1973): 215-81.
- Chesley, G.R., "The Elicitation of Subjective Probabilities: A Laboratory Study in an Accounting Context," Journal of Accounting Research, Spring 1976: 27-48.
- Cochran, W.G., "The Comparison of Percentages in Matched Samples," Biometrika, 37, 1950: 256-266.
- Collins, D., "Analytical Review: The Problem of Model Validity," Unpublished, University of Iowa, 1981.
- Craig, A., "Nonparametric Measures of Sensory Efficiency for Sustained Monitoring Tasks," Human Factors, 21, (1), 1979: 69-78.
- Daniel, T.C., L. Wheeler, R.S. Boster, and P.R. Best, "Quantitative Evaluation of Landscapes: An Application of Signal Detection Analysis to Forest Management Alternatives," Man-Environment Systems, 3, September 1973: 330-344.
- Deakin, E.B. and M.H. Granof, "Regression Analysis as a Means of Determining Audit Sample Size," The Accounting Review, October 1974: 764-771.

- Dawes, R.M. and B. Corrigan, "Linear Models in Decision Making," Psychological Review, 81, 1974: 95-106.
- Drury, C.G. and J.L. Addison, "An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance," Ergonomics, 16, 1973: 159-169.
- Egan, J.P., Signal Detection Theory and ROC Analysis. (New York: Academic Press, 1975).
- Egan, J.P. and F.R. Clarke, "Psychophysics and Signal Detection," in J.B. Sidowski (ed.), Experimentation Methods and Instrumentation in Psychology, (New York: Wiley, 1966).
- Egan, J.P., A.I. Schulman, and G.Z. Greenberg, "Operating Characteristics Determined by Binary Decisions and by Ratings," The Journal of Acoustical Society of America, 21, 1959: 768-773.
- Felix, W.L. and W.R. Kinney, "Research in the Auditor's Opinion Formulation Process: State of the Art," The Accounting Review, April 1982: 245-271.
- Ferrell, W.R., "Student Self-Assessment," Proceedings, Conference on Frontiers of Education, Tucson, 1972.
- Ferrell, W.R. and P.J. McGoe, "A Model of Calibration for Subjective Probabilities," Organizational Behavior and Human Performance, 26, 1980: 32-53.
- Glass, G.V. and J.C. Stanley, Statistical Methods in Education and Psychology, (Prentice Hall, 1970).
- Green, D.M., "Psychoacoustics and Detection Theory," The Journal of the Acoustical Society of America, 32, (10), 1960: 1189-1302.
- Green, D.M., "General Prediction Relating Yes-No and Forced-Choice Results," Journal of the Acoustical Society of America, 36, 1964: 1042.
- Green, D.M. and J.A. Swets, Signal Detection Theory and Psychophysics. (New York: Wiley, 1966).
- Grey, D.R. and B.J.T. Morgan, "Some Aspects of the ROC Curve-Fittings: Normal and Logistic Models," Journal of Mathematical Psychology, 9, 1972: 128-139.
- Hogarth, R.M., Judgment and Choice: The Psychology of Decision, (New York, John Wiley, 1980).

- Holder, W.W. and S. Collmer, "Analytical Review Procedures: New Relevance," The CPA Journal, November 1980: 29-35.
- Hosseini-Ardelahi, J., "Detectability of Correctness: A Signal Detection Measure of Knowing That One Knows," Unpublished Dissertation, University of Arizona, 1981.
- Howell, W.C. and S.A. Burnett, "Uncertainty Measurement: A Cognitive Taxonomy," Organizational Behavior and Human Performance, 22, 1978: 45-68.
- Hume, A., "Optimal Response Biases and The Slope of ROC Curves As a Function of Signal Intensity, Signal Probability, and Relative Payoff," Perception and Psychophysics, 2, 1974: 377-84.
- Hylas, R.E. and R.H. Ashton, "Audit Detection of Financial Statement Errors," The Accounting Review, October 1982: 751-65.
- Joyce, E.J., "Expert Judgment in Audit Program Planning," Studies on Human Information Processing in Accounting, Supplement to Journal of Accounting Research, 14, 1976: 29-60.
- Kadane, J.B. and S. Lichtenstein, "A Subjectivist View of Calibration," Decision Research Report 82-6, 1982.
- Kaplan, R.S., "Developing a Financial Planning Model for Analytic Review: A Feasibility Study," Third Symposium on Auditing Research, University of Illinois, 1978.
- Kinney, W.R. and A.D. Bailey, "Regression Analysis as a Means of Determining Audit Sample Size: A Comment," The Accounting Review, April 1976: 396-401.
- Kinney, W.R., "ARIMA and Regression in Analytical Review: An Empirical Test," The Accounting Review, January 1978: 48-60.
- Kinney, W.R. and G.L. Salamon, "The Effect of Measurement Error on Regression Results in Analytic Review," Symposium on Auditing Research III, University of Illinois, 1979.
- Kinney, W.R., "The Predictive Power of Limited Information in Preliminary Analytical Review: An Empirical Study," Journal of Accounting Research Supplement, 17, 1979: 148-165.
- Kinney, W.R. and W.L. Felix, "Analytical Review Techniques," Journal of Accountancy, October 1980: 98-103.

- Kinney, W.R., "Quantitative Applications in Auditing," Unpublished, University of Michigan, 1981.
- Lichtenstein, S. and B. Fischhoff, "Do Those Who Know More Also Know More About How Much They Know?", Organizational Behavior and Human Performance, 20, 1977: 159-183.
- Lichtenstein, S., B. Fischhoff, and L.D. Phillips, "Calibration of Probabilities: The State of the Art to 1980," in Kahneman, D., P. Slovic, and A. Tversky (eds.), Judgment Under Uncertainty: Heuristics and Biases, (New York: Cambridge University Press, 1982).
- Licklider, J.C.R., "Theory of Signal Detection," in Swets, J.A. (ed.), Signal Detection and Recognition by Human Observers, (New York: John Wiley, 1964).
- Luce, R.D., Individual Choice Behavior, (New York: Wiley, 1959).
- Lusted, L.B., "Signal Detectability and Medical Decision-Making," Science, 171, 1971: 1217-1219.
- Markowitz, J. and J.A. Swets, "Factors Affecting the Slope of Empirical ROC Curves: Comparison of Binary and Rating Responses," Perception and Psychophysics, 2, 1967: 91-7.
- McNemar, Q., Psychological Statistics, (New York: Wiley, 1949).
- McNicol, D., A Primer of Signal Detection Theory, (London: George Allen & Unwin Limited, 1972).
- Meehl, P.E., "When Shall We Use Our Heads Instead of the Formula?", Journal of Counseling Psychology, 4, 1957: 268-273.
- Mock, T.J., S.F. Biggs, and P.R. Watkins, "A Behavioral Framework for Evaluating the Use of Analytical Review Procedures in Auditing," University of North Carolina Conference on Audit Risk, May 1982.
- Mock, T.J. and P.R. Watkins, "Modelling Auditor Judgment," Symposium on Auditing Research IV, University of Illinois, 1980.
- Murphy, A.H., "A New Vector Partition of the Probability Score," Journal of Applied Psychology, 12, 1973: 595-600.
- Murphy, A.H. and R.L. Winkler, "Forecasters and Probability Forecasts: Some Current Problems," Bulletin of the American Meteorological Society, 52, 1971: 239-247.

- Nacmias, J., "Effects of Presentation Probability and Number of Response Alternatives on Simple Visual Detection," Perception and Psychophysics, 3, 1968: 151-55.
- Neter, J., "Two Case Studies on Use of Regression for Analytical Review," Center for Audit Research, Report 80-014, 1980.
- Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Brent, Statistical Package for The Social Sciences, (New York: McGraw-Hill, Inc., 1975).
- Ogilvie, J.C. and C.D. Creelman, "Maximum-Likelihood Estimation of Receiver Operating Characteristic Curve Parameters," Journal of Mathematical Psychology, 5, 1968: 377-91.
- Oskamp, S., "The Relationship of Clinical Experience and Training Methods to Several Criteria of Clinical Prediction," Psychological Monographs, 76, (574), 1962.
- Pastore, R.E. and C.J. Scheifer, "Signal Detection Theory: Considerations for General Application," Psychological Bulletin, 81, 1974: 945-958.
- Pease, K., R. Tarling, and P. Meudell, "Decisions in the Criminal Justice Process and Signal Detection Theory: A Note," Quality and Quantity, 11, 1977: 83-89.
- Pollack, I. and R. Hsieh, "Sampling Variability of the Area Under the ROC-Curve and of  $d'e$ ", Psychological Bulletin, 71, 1969: 161-73.
- Remus, W.E. and L.D. Jenicke, "Unit and Random Linear Models in Decision Making," Multivariate Behavioral Research, 13, 1978: 215-221.
- Robertson, J.C. Auditing, (Texas:: Business Publications, 1979).
- Schulman, A.I. and G.Z. Greenberg, "Operating Characteristics and A Priori Probability of the Signal," Perception and Psychophysics, 8, 1970: 317-20.
- Sheridan, T.B. and W.R. Ferrell, Man-Machine Systems: Information, Control, and Decision Models of Human Performance, (Cambridge, Mass: MIT Press, 1981).
- Shufford, H. and T.A. Brown, "Elicitation of Personal Probabilities and Their Assessment," Instructional Sciences, 4, 1975: 137-188.



- Smith, M. and W.R. Ferrell, "The Effect of Base Rate on Calibration of Subjective Probability for True-False Questions: Model and Experiment," Unpublished Working Paper (University of Arizona), 1981.
- Solomon, I., "Probability Assessment by Individual Auditors and Audit Teams: An Empirical Investigation," Journal of Accounting Research (forthcoming, 1982).
- Solomon, I., J.L. Krogstad, M.B. Romney and L.A. Tomassini, "Auditors' Prior Probability Distributions for Account Balances," Accounting, Organizations and Society, 7, 1982: 27-41.
- Statement of Auditing Standards, No. 1, (New York: AICPA, 1979).
- Statement of Auditing Standards, No. 23, Analytical Review Procedures, (New York: AICPA, 1981).
- Statement of Auditing Standards, No. 36, (New York: AICPA, 1981).
- Statement of Auditing Standards, No. 39, Audit Sampling, (New York: AICPA, 1981).
- Stringer, K.W., "A Statistical Technique for Analytical Review," Journal of Accounting Research Supplement, 1975: 1-9.
- Swets, J.A., "The Relative Operating Characteristic in Psychology," Science, 182, December 1973: 990-1000.
- Tanner, T.A., R.W. Haller and R.C. Atkinson, "Signal Recognition as Influenced by Presentation Schedules," Perception and Psychophysics, 2, 1967: 349-58.
- Taylor, D.H. and G.W. Giezen, Auditing: Integrated Concepts and Procedures, (New York: Wiley, 1979).
- Touche Ross & Co., Open Line, November 1981.
- Tukey, J.W., "Conclusions vs. Decisions," Technometrics, 2, 1960: 423-433.
- Wallace, W.A., "Discussants Response to 'The Effect of Measurement Error on Regression Results in Analytical Review'", Symposium on Auditing Research III, University of Illinois, 1979.
- Waller, W.S. and W.L. Felix, "The Auditor and Learning from Experience: Some Conjectures," Unpublished Working Paper, 1982.

Warren, C.S., "Discussion of 'A Statistical Technique for Analytical Review'", Journal of Accounting Research, Supplement, 1975: 10-13.

Yates, J.F., "External Correspondence: Decomposition of the Mean Probability Score," Organizational Behavior and Human Performance, 30, 198: 132-156.