CALIBRATION OF MEDICAL PROBABILITIES AT DIFFERENT LEVELS OF EXPERIENCE

by

Rita Yuk-King Huang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Administration and Curriculum
in Higher Education

1982

ABSTRACT

It is unclear whether medical training makes a difference in the accuracy and calibration of subjective probability, as it pertains to the incidence and prevalence of disease conditions. In an effort to contribute to knowledge in the area, the intent of this study was two-pronged: (1) to examine whether medical training leads to more accurate estimation of the incidence of acute and prevalence of chronic disease conditions; and (2) to determine whether medical training leads to better calibration.

Forty fourth-year medical students and forty non-medical students enrolled during the 1982 fall term at Michigan State University responded to a questionnaire by indicating their estimations of the relative incidence or prevalence of a series of paired acute and chronic diseases. They were also asked about their confidence in their estimations.

The results of the study showed that: (1) Medical students were significantly more accurate in their responses than non-medical students. This difference persisted even when the results were adjusted for age. (2) Medical students were also more confident of their responses than non-medical students. Age was not significantly related to

confidence. (3) Medical students were significantly better calibrated than non-medical students. When these results were adjusted for age, however, the differences became non-significant. Holding the accuracy scores constant, differences between the two groups in calibration persisted.

To my dearest husband, Raywin, our son Ritchie and our daughter, Rachelle, I dedicate this work.

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to many individuals who made the completion of this dissertation possible.

Special thanks to Dr. Arthur Elstein who has unselfishly given me his time and guidance throughout the process of this dissertation. He planted the idea for this study and his valuable critiques and suggestions on the proposal and the final research were extremely helpful. He unselfishly devoted much of his time to this study when I was working under severe time constraints. His humility, kindness and unselfishness serve as a model for me to follow.

I also wish to extend my thanks to Dr. R. Featherstone who allowed me to expand my research interests independently. Dr. H. Teitelbaum provided valuable critiques on the methodology of the study and Dr. N. Bell gave me his constructive suggestions on the proposal for the study. I appreciate their concern and guidance.

To my husband, Raywin, love, understanding, and patience served as emotional support throughout the whole process of this research, I owe a lifetime of gratitude. His advice on data analysis made the completion of my

thesis possible. My children, Ritchie and Rachelle, arrived at the beginning of my doctoral program and just as this study was nearing completion. They are delightful additions who made the completion of this work infinitely more challenging.

I also wish to express my love and gratitude to my parents who made my doctoral studies possible by providing initial support for my further education overseas.

To my Lord Jesus Christ whose encouragement and guidance helped me in times of frustration, I pledge eternal gratitude.

A final note of appreciation goes to the many friends who provided the help and kind words without which this total process of graduate study and research would have been cold, mechanical and futile. They have provided the laughter that made academic drudgery bearable and the caring and help that made complex problems solvable.

TABLE OF CONTENTS

		Page
LIST OF TABLES		iii
LIST OF FIGURES		iv
CHAPTER		
I.	STATEMENT OF THE PROBLEM	1
	Introduction	1
	The Problem	3
	Purposes	5
	Research Questions	6
	Research Hypotheses	6
	Definition of Important Terms	7
	Overview of the Study	9
II.	REVIEW OF LITERATURE	10
	Studies Related to Accuracy in Judgment	10
	Calibration in Judgment	13
III.	METHODS OF THE STUDY	21
	Research Design and Variables	21
	Research Variables	21
	Variable Matrix	22
	Research Procedures	24
	Population and Sampling	24
	Instrumentation and Data Collection	25
	Data Analysis	33

TABLE OF CONTENTS (Cont'd)

CHAPTER		Page
IV.	RESULTS, IMPLICATIONS AND RECOMMENDATIONS FOR FUTURE STUDY	34
	Results of the Study	34
	Discussion of the Findings	41
	Summary of the Findings	43
	Implications	44
	Recommendations for Future Study	47
BIBLIOGRAPHY		51
APPENDICES		54

LIST OF TABLES

Table		Page
I	Consistency (percent agreement) of the Duplicate Items in Parts I and II of this Instrument	32
II	Mean Scores, standard deviation and t-test differences in accuracy, confidence and calibration for medical and non-medical students . .	35
III(a)	Ages of Medical and Non-medical Students	37
III(b)	Correlation Between the Three Independent Measures: Accuracy, Confidence and Calibration	38
IV	Mean Subjective Probability for Medical and Non-medical Students .	39
V	Calibration: Analysis of Covariance Using Accuracy as Covariate	40

LIST OF FIGURES

Figure		Page
1.	Hypothetical Calibration Curve--A graph showing the percentages correct for each probability response . . .	4
2.	Mean Scores by Mean Subjective Probability for Each Subject in Medical and Non-medical Students	36

CHAPTER I

STATEMENT OF THE PROBLEM

Introduction

Making decisions is a daily human activity. The outcome of a decision may greatly affect the indivual's welfare or the welfare of others. Some decisions are based on personal belief such as voting decisions in a political election, policy decisions in business and deciding a trial verdict for a defendant in a courtroom. These beliefs are usually expressed in probabilistic language such as, "I think . . .," "chances are," "it is unlikely . . .," or "most probably . . .," and are usually based on intuition, knowledge of the event or subjective experience.

To improve accuracy in making decisions for a certain event, individuals have to know the actual probabilities of that event occurring and must "align" their beliefs with actuality. This process is known as the validation of subjective beliefs. One way of validation is by expressing these beliefs as estimates of subjective probability and comparing them with probability indexes mathematically derived from actual events. One example of a well-defined "actual" probability would be available reported actual

frequencies of events, such as rates of causes of death in the United States reported in Vital Health and Statistics. This technique of comparing a persons' subjective probability (a person's belief about the likelihood of an event) with the "actual" probability (the actual relative frequency of occurrence of the event) is called calibration. Individuals are perfectly calibrated if, over the long run, for all their given subjective probabilities, the proportion that is true is equal to the probability assigned (Fischhold, Slovic, Lichtenstein, 1977, p. 522). For example, a perfectly calibrated individual in assigning events a probability of .7 will have 70 percent of those responses correct and for all responses assigned a probability of .8 will be 80 percent correct.

However, people are not always perfectly calibrated in their probability estimations. Several research studies, reviewed by Lichtenstein, Fischhoff, and Phillips (1977), have shown that people tend to be biased in their probability estimations. In other words, people tend to over or under estimate how much they know. For individuals who are underconfident the proportion of responses that are correct is greater than the probability assigned to them (Lichtenstein and Phillips, 1977, p. 276). For example, individuals might be only 50 percent certain of their responses but have 90 percent of those responses correct.

They underestimate how much they know. In cases of overconfidence the proportion of correct responses is less than the probability assigned to them. For example, people might be 90 percent confident in their responses, but have only 75 percent of those responses correct. That is, they overestimate how much they know and believe they know more than they actually do. A graph showing the percentage correct for each probability response is plotted in Figure 1. Curve A reflects underconfidence, curve B represents perfect calibration and curve C represents overconfidence.

From this discussion, two important underlying dimensions are noted. They are: (1) accuracy of estimation (that is, percent correct in responses); and (2) the degree of confidence placed in an estimation. What makes individuals better calibrated in their estimations? Do training and experience lead to more accurate and better calibration in estimation?

The Problem

Tremendous amounts of time, energy and resources, especially at higher levels of learning in higher education, are spent to attain high levels of expertise needed in various areas of specialization. The expected outcome of specialization is that the so-called experts will become better decision-makers in the areas in which they have received

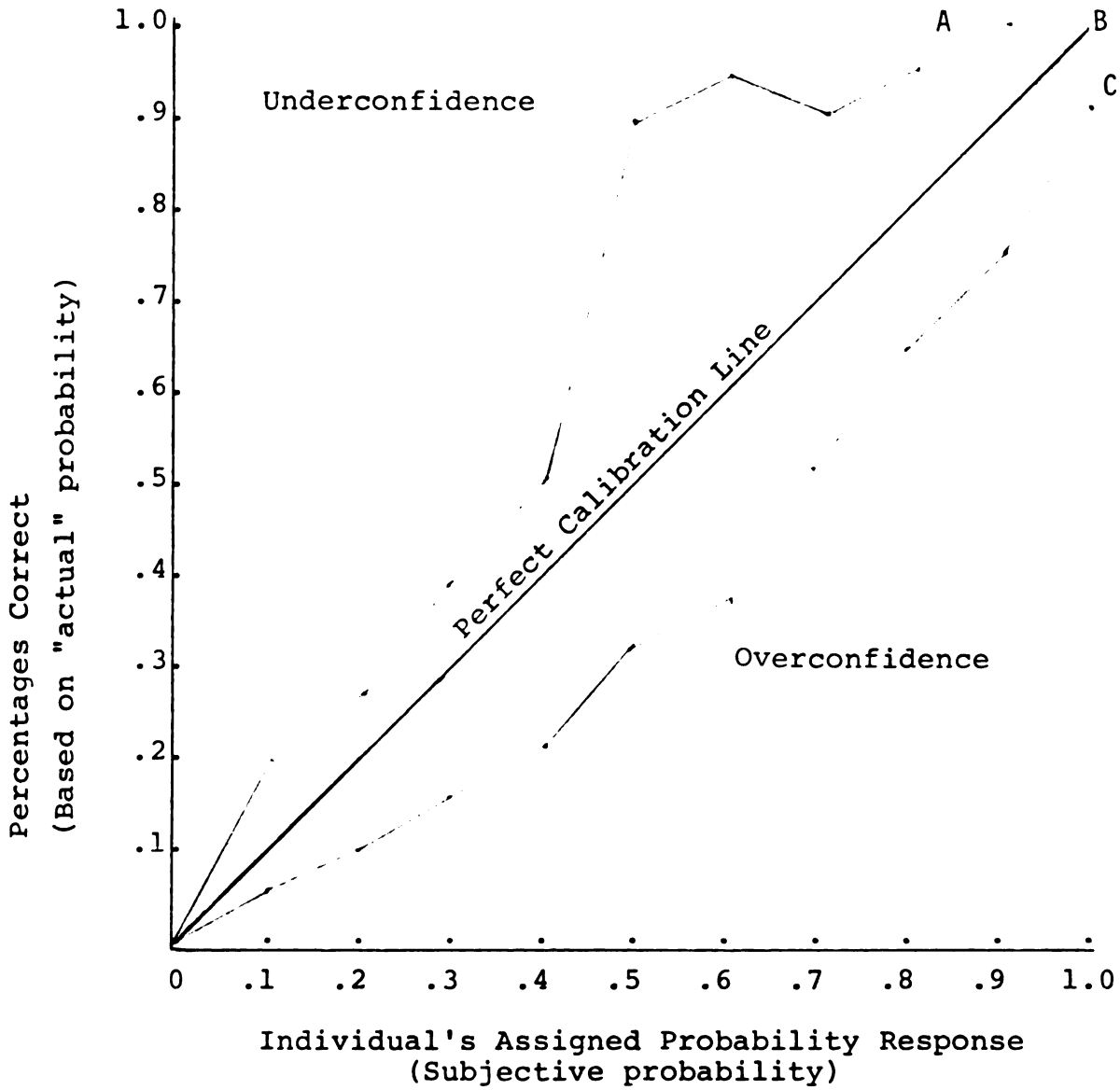

Figure 1. Hypothetical Calibration Curve--A graph showing the percentages correct for each probability response

training, than those who do not receive the same training. This presupposition acts as the impetus for this study to re-examine the notion that knowledge and training lead to better decision-makers who are more accurate and better calibrated on probability estimates in their areas of specialization.

Whether training and knowledge will really make a difference in improving the accuracy and calibration of subjective probabilities is unclear. The inconsistent results of previous research are presented in the literature review section of this study. Such inconsistent results may be due to the short-term nature of the training tested in the studies. The training frequently had little time to take effect to improve the subjects' judgment performances. Hence, the present research was undertaken to examine a longer term of training, such as exists in higher education where training is systematically planned and structured toward a defined career goal. The training in higher education is also more intensive, with frequent problem-solving exercises and examinations. Thus, such training should have significant effect in improving trainee's capability for making accurate decisions and building confidence in those decisions, thereby leading to better calibration.

Purposes

This study has two purposes:

(1) To reexamine the question of whether medical training leads to more accurate estimates of the incidence of acute and prevalence of chronic disease conditions.

(2) To determine whether medical training leads to better calibration.

Research Questions

The following questions summarize the two central issues of this study: accuracy and calibration.

Accuracy

(1) Do people who have medical training know more about the incidence of acute and prevalence of chronic disease conditions, than people who do not have medical training?

Calibration

(2) If they do, are people with medical training (experts) calibrated better than people without medical training (non-experts)?

Research Hypotheses

The following hypotheses were formed to examine the questions posed:

(1) Medical students are significantly more accurate than non-medical students in estimating the incidence of acute and prevalence of chronic disease conditions.

(2) Medical students are significantly better calibrated than non-medical students in estimating the incidence of acute and prevalence of chronic disease conditions.

The testing of these hypotheses will provide empirical evidence bearing on the research questions.

Definition of Important Terms

The following definitions for key terms used in the study will serve to provide a common basis for understanding.

-- Medical students. Fourth year medical students enrolled in the College of Human Medicine at Michigan State University during the Fall Term of the academic year 1981-1982.

-- Non-medical students. Michigan State University senior or graduate students who have a non-medical major (excludes those majoring in nursing, medical technology, osteopathic medicine, veterinary medicine, and other health-related areas). These students were also enrolled in the Fall Term, 1981.

-- Accuracy. Percentages of items correctly identified by the subject. The correct answers to the questions were derived from the statistics in Vital and Health Statistics, Series 10, Nos. 109 and 132.

-- Incidence of acute disease conditions. New cases of acute disease conditions occurring among people in the United States based on statistics derived from series 10,

No. 132, of Vital and Health Statistics. The acute conditions selected for this investigation were:

Influenza	Fracture and dislocation
Pneumonia	Sprain and strain
Headache	

-- Prevalence of chronic disease conditions. The cases (including new and old cases) of chronic disease conditions existing among people in the United States. The statistics were derived from series 10, No. 109, of Vital and Health Statistics. The chronic disease conditions included in this study were:

Arthritis	Epilepsy
Cerebrovascular disease	Heart conditions
Diabetes	Tuberculosis

-- Better calibration. A person is perfectly calibrated if, over the long run, for all responses assigned the same probability, the proportion correct is equal to the probability assigned. A perfect calibration score is 0 and the worst possible score is 1.0.

-- Confidence. A person who is not perfectly calibrated can be either overconfident or underconfident. A person is overconfident when the portion of responses that are correct is less than the probability assigned to them. Overconfidence is shown by a positive score. A person is

underconfident if the portion of correct responses is greater than the probability assigned to them. Underconfidence is shown by a negative score.

Overview of the Study

In Chapter I the problem, research questions and hypotheses have been stated. Important terms have been defined. In the chapter to follow a review of research studies related to calibration will be presented. Chapter III will include the research design and variables and will focus on research procedures. In Chapter IV the results of the study will be presented and discussed. The findings will be summarized and their implications considered. Recommendations for future study will be outlined.

CHAPTER II

REVIEW OF LITERATURE

In this chapter a review of research studies related to calibration are presented. The review of literature is organized under two major headings: (1) Studies related to accuracy in judgment; and (2) studies related to calibration in judgment. Accuracy measures the percentage of items answered correctly by respondents. Calibration measures confidence in a subject's judgment.

Accuracy in Judgment

In this study accuracy in judgment is determined by comparing the subjects' chosen answers with a criterion. Accuracy is measured by percent correct.

Studies have shown that subjects with no training or with no prior knowledge in a particular area tend to have difficulties performing a task in that area. In an early study, Lichtenstein and Fischhoff (1976) asked subjects with a limited knowledge of painting to study small sketches drawn by European and Asian children and determine if the artist was an European child or an Asian child. Results showed that the subjects had difficulty with the task. Only 53.2 percent of their 1,104 answers were correct.

In another experiment, reported in the same study, other subjects with limited knowledge of stocks were asked to study market charts and predict whether a stock described in each chart would be up or down three weeks later. This task was even more difficult for subjects to perform accurately and only 47.2 percent of their choices were correct.

These studies concluded, therefore, that subjects without training experience or knowledge in a particular area tend to have difficulty in performing judgment tasks in that area.

Training and knowledge, however, might be assumed to affect subjects' performance in terms of accuracy. Another experiment done by Lichtenstein and Fischhoff (1976) required subjects to identify handwritings, determining whether they were written by an European or an American. Two of the four groups received training for this task. Results showed that trained subjects correctly identified 71.4 percent of the specimens compared with 51.2 percent for untrained subjects.

The question of whether there will be a difference in performance in terms of accuracy for subjects with high levels of knowledge or experience in a certain task is discussed in several studies with conflicting results. Sanders (1963) found that students and instructors of meteorology

performed about equally well in weather forecasting. Gustafson (1963) compared diagnoses of congential heart disease made by a computer, pediatric cardiologists, and non-specialized physicians. The pediatric cardiologists and the computer appeared to be about equally accurate, correctly diagnosing 63-74 percent of the cases, while the non-specialized physicians were less accurate, correctly diagnosing only 36-52 percent of the cases. Another study conducted by Gustafson (1966), however, found basically no differences between surgical residents and experienced surgeons in their ability to predict patients' length of stay in the hospital.

Winkler (1967, 1971) has shown that being an expert in one's own field leads to better performance. He found that sportswriters and bookmakers were better than college students and faculty at predicting scores of NFL and Big Ten football games.

Stael Van Holstein (1971) compared four groups of subjects in forecasting the weather. The four groups of subjects with different levels of knowledge in meteorology were meteorologists, meteorology research assistants, students of meteorology and statisticians. It was found that the research assistants showed the best forecasting ability, the students the worse, the meteorologists' and statisticians' forecasting ability fell in between, indicating a curvilinear relationship between level of expertise and

accuracy of judgment.

Stael Van Holstein's (1972) experiment compared the performance of five groups of people - bankers, stock market experts, statisticians, teachers of business administration and students of business administration in predicting the variability of the stock market. Results showed that stock market experts and statisticians performed the best, followed by business teachers and students, and bankers were last.

In summary, previous research has demonstrated that subjects with no prior training or knowledge in a specialized area tend to have difficulties making accurate judgments in the area. However, subjects with even a minimal amount of knowledge or with a minimum of prior training improve slightly in performance accuracy. There is not much evidence to determine whether or not expertise in a speacialized area increases accuracy in judgments (Beach, 1975). Some studies (Winkler, 1967, 1971; Stael Van Holstein, 1972) reported that experts in a particular field tend to have better performances in that field. Other studies (Gustafson, 1963, 1966; Sanders, 1963) concluded that non-experts performed equally well in specialized areas.

Calibration in Judgment

Measuring accuracy of judgment, however, does not

capture the certainty or the confidence subjects have in their judgments. A subject can achieve good calibration with little or no knowledge or experience in an area. On the other hand, a person with experience and knowledge in a particular area may not achieve good calibration; the person may believe he knows more than he actually knows (overconfidence) or less than he actually knows (underconfidence).

At least one published study found that subjects who are not experts in a particular area may achieve good calibration. Using full-range approach with four alternative items, Fischhoff and Beyth (1975) asked 150 Israeli university students who were not foreign affair experts to assess the probability of 15 then-future events, such as "President Nixon will meet Mao at least once." The resulting calibration curve is suboptimal at 0 and 1, and shows a dip at .7 but is otherwise remarkably close to the identity line (perfect calibration).

Other studies show that subjects with no training and no prior knowledge in a particular area were poorly calibrated. In fact, they showed no evidence of calibration at all. In one of their experiments discussed earlier, Lichtenstein and Fischhoff (1976) asked subjects to identify small sketches drawn by European and Asian children. Results indicated that subjects were overconfident in their judgments. In another experiment in the same study, subjects studied stock market charts and were asked to predict whether the stock prices described by each chart would be increasing or decreasing. Again, subjects demonstrated

overconfidence. The above two experiments demonstrated that subjects with no prior knowledge or training in an area tend to be overconfident in their calibration, in contrast to the almost perfect calibration obtained in the 1975 Fischhoff and Beyth study.

In an effort to determine whether training would improve calibration, Adams and Adams (1958) asked subjects to decide whether pairs of words were antonyms, synonyms, or unrelated. Calibration tallies and calibration curves were shown to subjects after each of five training sessions as feedback. A modest improvement in calibration was found after subjects received training.

In another study by Pratt (1977) an expert was asked to predict attendance at 175 movies shown in local theaters over a period of more than one year. This expert was given some degree of additional training by receiving feedback throughout the experiment. Results indicated that the only evidence of improvement in calibration over time came in the first few days; no further improvement was noted later.

Pickhardt and Wallace (1974) also reported slight improvement in calibration with five or six training sessions on estimation of uncertain quantities. However, in another study done by the same researchers, using a simulation game called PROSIM that was intended to increase know-

ledge on calibration, increased information did not affect calibration. There was virtually no improvement in calibration over the nineteen days of simulation. Using only one training session on 75 items, Chou (1976) also found little improvement and no generalization in calibration.

Lichtenstein and Fischhoff (1976) asked two groups of subjects to examine ten handwritings to determine whether they had been written by an European or an American. The training group studied samples of handwritings labeled with country of origin, the non-training group studied samples of handwriting that were unlabeled. Results showed that trained subjects showed better calibration than untrained subjects, who showed no evidence of calibration.

Lichtenstein and Fischhoff (1980) trained people without previous experience in probability assessment using computerized feedback provided after sessions of assessment on general knowledge items. Eleven long and intensive sessions were used. Results showed that most subjects' calibration improved during the training sessions. The mean calibration score changed from .015 for the first session to .005 for the last session. All measurable improvements were found to come between the first and the second rounds of training.

The above studies demonstrated that calibration can be somewhat improved by training. All the above studies,

however, were done using short-term training. No studies of prolonged training related to calibration were found.

Studies of subjects with different levels of knowledge, experience and training (experts vs. non-experts) were examined to determine whether the subjects would be calibrated differently. Oskamp (1962) divided subjects into three groups with varying levels of experience to evaluate the MMPI profile. The three groups were:

(1) 28 undergraduate psychology students representing inexperienced judges.

(2) 23 clinical psychology trainees working at a VA hospital; and

(3) 21 experienced clinical psychologists.

Their task was to determine whether VA hospital patients had been admitted for psychological reasons or medical reasons simply by reviewing their MMPI profiles. The subjects were then asked to assign a probability of correctness to their decisions in each case. Results showed that all three groups were overconfident, especially the undergraduates in their first session. When the first group was split into two groups, one with training for accuracy and the other without training, the trained groups showed better calibration.

Sanders (1963) asked students of meteorology and instructors of meteorology to forecast the weather and

found that students tend to overestimate the probability that an event will occur. Hazard and Peterson (1973) asked 40 subjects at the Defense Intelligence School to respond to 50 two-alternative general knowledge items. Substantial overconfidence was also found in this study.

Lichtenstein and Fischhoff (1976) in their experiment asked 120 subjects to answer general knowledge items. Based on the accuracy of their responses, the subjects were divided into three subgroups according to their knowledge: the best subjects (40 subjects with 51 or more correct answers out of 75), the middle subjects (39 subjects with 46-50 correct answers, and the worst subjects (41 subjects with fewer than 46 correct answers). Separate analyses were performed for each group. The result showed that subjects' calibrations varied directly with their knowledge. All groups tended to be overconfident. The most knowlegeable subjects showed the least overconfidence and had a calibration curve closest to the identity line. The results strongly suggested that the more subjects know, the better their calibrations are.

In another experiment, Lichtenstein and Fischhoff asked graduate students in psychology to answer 50 general knowledge items and 50 specially written items dealing with psychology. The two types of items were intermixed randomly in the stimulus package. The subjects were split for analysis, into best and worst at the median (74.5%)

of the distribution of percentage correct. The items were also split into easy (at least 75 percent correct) and hard (fewer than 75 percent correct) items. For these analyses, no distinction was made between general knowledge and psychology items. Results showed that the group with the greatest knowledge (best subjects in terms of percentage correct) did not have the best calibration scores. The most knowledgeable subjects in answering the easiest items showed substantial underconfidence, while the worst subjects, in responding to the hardest items, showed substantial overconfidence. In another experiment Lichtenstein and Fischhoff asked subjects with different levels of knowledge to answer randomly intermixed questions with 50 general items and 50 psychology items. Here again, results showed that the group with the greatest knowledge did not have the best calibration score.

The above calibration studies demonstrate that people are prone to systematic biases in their probability judgments. The most common bias is overconfidence because they believe that they know more than they actually know (Lichtenstein and Fischhoff, 1980, p. 2). Another conclusion is that training can sometimes improve calibration. And finally, people who have knowledge in a

specialized area (experts) sometimes demonstrated better calibration and sometimes not. Previous studies on subjects' accuracy in judgment also demonstrated that training and knowledge can improve accuracy, at least for a short period of time.

CHAPTER III

METHODS OF THE STUDY

In this chapter the research design and research procedures of the study will be discussed in separate sections.

Research Design and Variables

(1) Research Variables

The independent variable of this study was level of medical knowledge represented by two groups, medical and non-medical students. There were three dependent variables: accuracy, level of confidence, and calibration.

(a) Accuracy - is measured by percent of correct responses in identifying the incidence of acute and prevalence of chronic conditions. It can be expressed as:

$$\text{Accuracy} = \frac{n}{N}$$

Where n is the number of items correctly identified and N is the total number of items.

(b) Confidence (over/underconfidence) - the subject's level of confidence in making a decision. It is

defined as follows:

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

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 correct for all times assigned probability r_t, and T is the total number of different response categories used. Overconfidence is shown by a positive difference and underconfidence by a negative difference.

(c) Calibration - this measure, derived from Murphy (1973), is:

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

A perfect calibration would have a value of 0 and the worst possible score would be 1.0 which can be obtained by a subject who always responds $r_t = 1$ when wrong, and $r_t = 0.0$ when right.

(2) Variable Matrix

Given the above mentioned independent variable and dependent variables, a variable matrix can be drawn as follows:


Levels of specialization

Calibration Measures

Levels of Specialization	Accuracy	Confidence	Calibration
Medical Students			
Non-Medical Students			

Simple independent t-tests are used to test the different groups. The critical significance level of < 0.05 is used to test all hypotheses.

Research Procedures

(1) Population and Sampling

(a) Population

The subjects of this study were divided into two groups representing two different levels of training. They were:

(i) Medical students--These students were in their fourth year of medical training and are assumed to possess a certain level of medical knowledge.

(ii) Non-medical students--These students were seniors and graduate students studying in fields other than medicine or health related areas. They presumably possess a minimum level of medical knowledge.

(b) Sampling

Forty fourth-year medical students from Michigan State University voluntarily participated

in this study and were used as the subjects for the first group. These 40 fourth-year medical students were enrolled in the College of Human Medicine and do not include students majoring in veterinary medicine, osteopathic medicine, nursing and other health-related areas.

Forty fourth-year or graduate students in major areas of study other than medicine or other health-related fields volunteered to participate in the study and were used as subjects for the second group (non-medical students). These 40 non-medical Michigan State University students were majoring in such fields as business, education, forestry, mathematics, theatre, human ecology, engineering, computer science, psychology, sociology, audiology and communication.

Only students who had been in the United States more than ten years were used as subjects in either groups. The underlying reason was to exclude those students from other countries who might not be familiar with the subject matter of this study.

(2) Instrumentation and Data Collection

(a) Instrumentation

A questionnaire was designed as the assessment instrument for the study. The format of the instrument

is comprised of pairs of conditions (or diseases) presented to the subjects. Each question presents two alternative answers, one of which is true, the other false. Subjects are asked to identify which alternative is true, and to indicate the probability that the chosen alternative is, in fact, true. A sample of the instrument items is presented in Appendix A.

The procedure for selecting the various acute and chronic conditions used in Part I and Part II of the questionnaire is described below.

Part I: Acute Conditions

Six acute conditions were selected from Table 1, Incidence of Acute Conditions, Percent Distribution, and Number of Acute Conditions per 100 Persons per Year, by Condition Group, According to Sex: United States, July 1977 - June 1978, appearing on pp.11-12 of National Vital and Health Statistics, Series 10, No. 132. The procedure for selecting these six acute conditions was as follows:

(1) Three categories of conditions (respiratory conditions, injuries and other acute conditions) that had the highest frequencies of occurrence were selected from the five categories of conditions presented.

(2) Within each of the three categories selected, conditions were rank ordered by frequency of occurrence from highest to lowest. The median of the ranking was used to divide the high frequency from the low frequency group.

(3) Then, one acute condition was randomly selected from the high frequency group and one acute condition was selected from the low frequency group in each of the three categories of conditions.

(4) From respiratory conditions, "influenza" was selected from the high frequency group and "pneumonia" was selected from the low frequency group.

(5) From the category of injuries, the high frequency acute condition selected was "fractures and dislocations" and the low frequency acute condition selected was "sprains and strains."

(6) From the category under "other acute conditions," "diseases of the ear" was selected from the high frequency group and "headache" was selected from the low frequency group.

(7) The six selected acute conditions: influenza pneumonia, fractures and dislocations, sprains and strains, diseases of the ear, and headache were then arranged in alphabetical order and assigned a number from 1 to 6. Fifteen possible pairs of acute conditions can be formed from the six acute conditions chosen (See Appendix A).

(8) These 15 pairs of acute conditions compromise the 15 items in Part I of the questionnaire.

(9) Five items were randomly selected to check for reliability. By reversing the order of the disease conditions for five pairs--(2,1), (3,2), (4,3), (5,4), (6,5)--items 16 to 20 were formulated in Part I of the questionnaire.

Part II. Chronic Conditions

In the second part of the questionnaire, six chronic conditions were selected from "Table K, Number per 1,000 Persons, Prevalence, and Incidence of Selected Chronic

Conditions Reported in Health Interviews: United States, 1968-1973" appearing in Vital and Health Statistics, Series 10, No. 109. The procedures for selecting these six chronic conditions are described below:

(1) Since the prevalence of chronic conditions in Table K were arranged from highest frequency to the lowest, the prevalence rate of 10.3 per 1,000 persons was used as the point at which to divide the conditions into high and low frequency groups.

(2) Three chronic conditions were randomly selected from the high frequency group and three other chronic conditions were selected from the low frequency group.

(3) The three chronic conditions selected from the high frequency group were arthritis, diabetes and heart conditions. The three chronic conditions selected from the low frequency group were cerebrovascular disease, epilepsy and tuberculosis.

(4) These six chronic conditions were arranged in alphabetical order and each was assigned a number from 1 to 6. Fifteen possible pairs of chronic conditions can be formed from the six chronic conditions chosen (See Appendix B).

(5) The 15 pairs of chronic conditions comprise the 15 items in Part II of the questionnaire.

(6) Five items were randomly selected to check reliability. By reversing the order of the disease conditions, that is, (2,1), (3,2), (4,3), (5,4), items 16 to 20 were formulated for Part II of the questionnaire.

The subjects were asked to state how confident they were about their chosen answer by circling a confidence

rating from 0% to 100% under each item in both Part I and Part II of the questionnaire. An example of such an item is shown below.

A. Headache

B. Influenza

How confident are you that your answer is correct?

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

A glossary of terms were also attached to the instrument (Cooley, 1973).

(b) Data Collection

The instrument is presented in Appendix C. It was administered to three medical students and five non-medial students as a pilot study to determine the reliability of the instrument. It will be remembered from the discussion of the instrument that five duplicate items with the order of the disease conditions reversed were added to both Part I and Part II to serve as a check of reliability. The five pairs of duplicate items (Nos. 16 to

20) in Part I are numbers 2 and 16, 3 and 17, 10 and 18, 13 and 19, 15 and 20. The five pairs of duplicate items (Nos. 16 to 20) in Part II are numbers 13 and 16, 10 and 17, 5 and 18, 12 and 19, 4 and 20.

Percent agreement (that is, the degree of consistency in selecting the same answer for the pairs of the duplicate items) was calculated for each pair of duplicate items. An overall percent agreement on the duplicate items was also calculated separately for Parts I and II. The overall consistency of the pairs of duplicate items in Part I was 87.8 percent agreement. In Part II, the overall consistency of the pairs of duplicate items was 85.2 percent (See Table I).

From this pilot study the conclusion can be drawn that the instrument to be used for assessing calibration is highly reliable, indicating that subjects were not responding randomly.

Upon completion of the pilot study, the instrument was administered to 40 non-medical student volunteers from various majors as mentioned earlier. Ninety-five questionnaires were then mailed to five medical training communities affiliated with Michigan State University in Kalamazoo, Saginaw, Lansing, Flint and Grand Rapids, where fourth-year medical students have clerkships. Administrators in charge of the medical training communities were

Table I. Consistency (percent agreement) of the Duplicate Items in Parts I and II of the Instrument

Part I		Part II	
Pairs of duplicate items with order of conditions reversed	% of Agreement	Pairs of duplicate items with order of conditions reversed	% of Agreement
Nos. 2 and 16	100	Nos. 13 and 16	88
Nos. 3 and 17	88	Nos. 10 and 17	75
Nos.10 and 18	88	Nos. 5 and 18	100
Nos.13 and 19	88	Nos. 12 and 19	75
Nos.15 and 20	75	Nos. 4 and 20	88
Overall Mean Percent Agreement	87.8	Overall Mean Percent Agreement	85.2

18, 13 and 19, 15 and 20. The five pairs of duplicate items (Nos. 16 to 20) in Part II are numbers 13 and 16, 10 and 17, 5 and 18, 12 and 19, 4 and 20.

Percent agreement (that is, the degree of consistency in selecting the same answer for the pairs of the duplicate items) was calculated for each pair of duplicate items. An overall percent agreement on the duplicate items was also calculated separately for Parts I and II. The overall consistency of the pairs of duplicate items in Part I was found to be 87.8 percent agreement. In Part II, the overall consistency of the pairs of duplicate items was 85.2 percent (See Table I).

From this pilot study the conclusion can be drawn that the instrument to be used for assessing calibration is highly reliable, indicating that subjects are not randomly choosing the responses.

Upon completion of the pilot study, the instrument was administered to 40 non-medical student volunteers from various majors as mentioned earlier. Ninety-five questionnaires were then mailed to five medical training communities affiliated with Michigan State University in Kalamazoo, Saginaw, Lansing, Flint and Grand Rapids, where fourth-year medical students have clerkships. Administrators in charge of the medical training communities were asked to distribute the questionnaires to the students in

asked to distribute the questionnaires to the students in each community. Forty completed questionnaires were returned and were used as data for the group of medical students.

(3) Data Analysis

The data from the groups of non-medical and medical students were entered onto a master computer file and analyzed using the Michigan State University CDC cyber 750, with the Statistical Package for Social Science program (SPSS). Independent t-tests were used in testing the difference between the groups for accuracy, over/underconfidence and calibration scores. In the chapter to follow the results and findings from this analysis will be discussed.

CHAPTER IV

RESULTS, IMPLICATIONS AND RECOMMENDATIONS FOR FUTURE STUDY

In this chapter the findings of the study, the results and the implications will be discussed. Recommendations for future studies will be presented.

Results of the Study

The derivation of the three independent measures accuracy, confidence and calibration scores and an example of their computation are presented in detail in Appendix D. The mean scores, standard deviation and t-test differences of the three measures are summarized in the variable matrix in Table II. The difference between the two groups is more obvious when the mean scores for each individual on accuracy (percent correct) and mean subjective probability are plotted in relation to the calibration curve, as shown in Figure 2.

The results of each measure: accuracy, confidence and calibration will be presented separately in the following paragraphs.

Accuracy. Table II gives the mean score, standard deviation and the t-test differences of accuracy for all

Table II. Mean Scores, standard deviation and t-test differences in accuracy, confidence and calibration for medical and non-medical students

Groups	Accuracy			Confidence[1]			Calibration[2]		
	M	SD	Observed t	M	SD	Observed t	M	SD	Observed t
Medical Students (N=40)	76.9%	8.8	3.22**	0	.13	2.11*	.07	.04	2.06*
Non-medical Students (N=40)	71.2%	7.1		-.09	.23		.12	.13	

[1]Negative value indicates underconfidence

[2]The best possible calibration score = 0; and worst = 1.0

df = 78

*p = .04

**p = .002

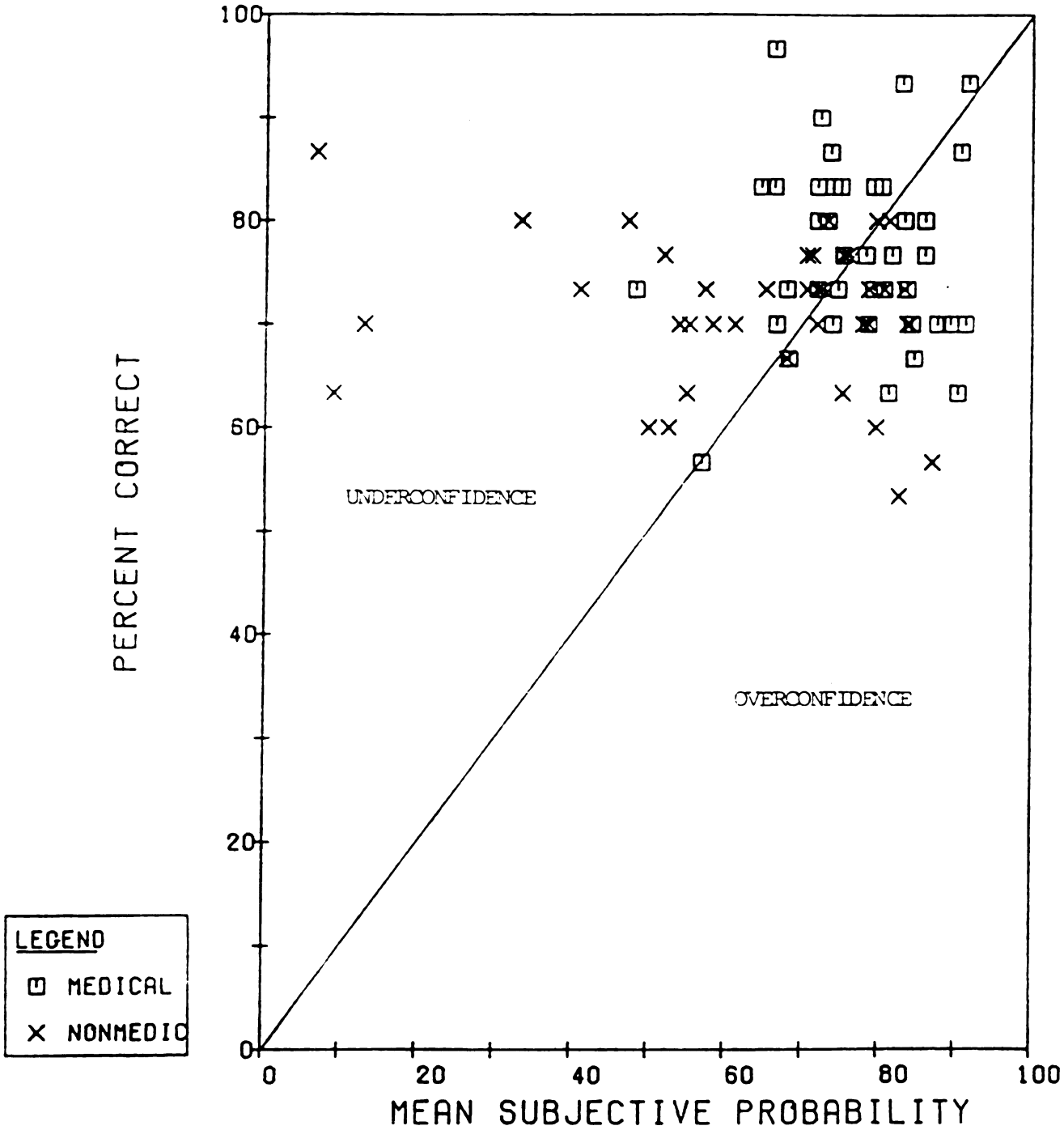

Figure 2. Mean Scores by Mean Subjective Probability for Each Subject, in Medical and Non-medical students

medical students and non-medical students. Statistically, there is a significant difference (p = .002) between medical students and non-medical students on the accuracy measure. The mean accuracy score showed that medical students correctly identified 76.9 percent of the disease conditions correctly, compared with 71.2 percent for non-medical students.

Since many of the subjects in the non-medical group were fourth-year undergraduate students and younger than the medical students (See Table III) an analysis of co-variance was used to adjust for age differences. It was found that by adjusting the age, there was still a significant difference (p = .001) between medical and non-medical students. Age, therefore, was not a factor in determining accuracy. Accuracy was determined more by specific training than by general experience.

Table III (a). Ages of Medical and Non-Medical Students

Group	Age		
	M	SD	p
Medical Students	27.8	3.3	.04
Non-medical Students	23.6	4.6	

Table III (b). Correlation Between the Three Independent Measures: Accuracy, Confidence and Calibration

	Age	Accuracy	Confidence	Calibration
Age		.21 p=.03	.01 p=.47	-.18 p=.06
Accuracy			-.38 p=.001	-.05 p=.34
Confidence				-.63 p=.001

From the above analysis, it may be concluded that medical students are more accurate than non-medical students in estimating the incidence of acute and the prevalence of chronic disease conditions even though the difference is not large (i.e. 5-6 percent difference). This difference persists even when scores are adjusted for age. It is also worth noting that all the subjects achieved more than 50 percent accuracy. A possible explanation of these results will be discussed in a later section.

Confidence. There was a significant difference ($p = .04$) between medical and non-medical students on the confidence measure, as shown in Table II. Medical students who had a mean subjective probability of 77 percent were more certain of their responses than non-medical students

who had a mean subjective probability of 63 percent (See Table IV).

Table IV. Mean Subjective Probability for Medical and Non-medical Students

	Mean Subjective Probability	
	M	SD
Medical students	77%	9.5
Non-medical students	63%	20.3

Table II also showed that non-medical students were underconfident in their estimations with a confidence score of -.09. Medical students were neither overconfident nor underconfident (confidence score = 0). Figure 2 shows that 62 percent of the subjects in the non-medical group placed in the underconfidence side of the graph and that 38 percent of the subjects lie on the overconfidence side of the graph. Medical students' estimates of their own responses were equally divided between overconfidence or underconfidence.

Since confidence and age were not correlated (as shown in Table III (b), age was not used as a covariate for adjusting confidence measures.

The above analysis has concluded that there was a significant difference between the groups of subjects (40 medical and 40 non-medical students) in confidence and that this difference is unrelated to the age difference between the groups. Non-medical students showed underconfidence in their estimation whereas medical students tended to be neither underconfident nor overconfident.

Calibration. In Table II there was a significant difference ($p = .04$) between medical and non-medical students on calibration. Medical students, with a calibration score of .07, were closer to perfect calibration of "zero" and non-medical students had a calibration score of .12.

Using covariance analysis to adjust the age differences between the two groups, it was found that there were no significant differences ($p = .14$) between medical and non-medical students in calibration. Using the accuracy score as a covariate, the differences ($p = .05$) in calibration persisted (See Table V).

Table V. Calibration: Analysis of Covariance Using Accuracy as Covariate

	Sum of Squares	df	Mean Square	f	Significance of f
Group Adjusted	.04	1	.04	4.4	.05
Within Group Adjusted	.72	77	.01		
Total		78			

To summarize the results of the calibration measure, it may be concluded that students with medical training were better calibrated than students without medical training. Adjusting for age, there was no significant differences between the calibration scores of the groups. Using the accuracy score as covariate, the difference in calibration persists.

Discussion of the Result

Although there was a significant difference between medical and non-medical students in accuracy, that difference was very small (5-6 percent). It was also found that all subjects achieved at least 50 percent accuracy. Two possible explanations for the above findings will be discussed.

One underlying factor might be the format of the items. Since there were only two choices for each item, the respondent had a 50-50 chance of being correct (or incorrect). Even if the respondents did not know the answers, they could guess and still have a 50 percent chance of being correct. This fact may explain why the differences in accuracy between medical and non-medical students were so small. It is also apparent that all subjects had a tendency to obtain high accuracy scores due to the two-choice format of the items.

A second possible explanation is that the task itself

might be composed of a certain percentage of general medical knowledge with the rest representing specialized medical knowledge. General medical knowledge can be obtained from public media (newspaper reports, laymen magazines, etc.) whereas the specialized medical knowledge could only be obtained from formal education. If this explanation is true, one can estimate the percentages from the performances of the groups. In this task, all non-medical students achieved an average 70 percent accuracy, therefore, one can assume that there are approximately 70 percent general medical knowledge items and the other 30 percent (i.e. 100%-70%=30%) are specialized medical knowledge items. In this 30 percent of specialized medical knowledge items, 6 percent represent medical knowledge obtained from formal training, as the difference between medical students and non-medical students in accuracy is approximately 6 percent (i.e. 76%-70%=6%). The balance of the 30 percent or 24 percent (i.e. 30%-6%=24%) represent even more specialized medical knowledge which cannot be obtained from formal training but are only learned from longer experience in the medical field.

The variable, age, implies training and general experience. We know that medical students are more accurate than non-medical students simply because of training. Does training, then, also aid calibration? To test this an analysis of covariance was then applied using accuracy as the

covariate and it was found that the difference in calibration between medical and non-medical students retains its significance. When age was used as a covariate, the difference between the two groups in calibration were no longer significant. Hence, it may be general experience, rather than training, contributes to calibration.

Summary of the Findings

The major findings of the study can be summarized as follows:

(1) There was a significant difference between students with medical training and students without medical training on accuracy in estimating the incidence of acute and prevalence of chronic disease conditions. This difference persists even when adjusted for age. Thus, medical students were slightly more accurate than non-medical students in estimating the incidence of acute and the prevalence of chronic disease conditions.

(2) There was a significant difference between students with medical training and students without medical training on confidence in estimating the incidence of acute and prevalence of chronic disease conditions. Age was not significantly related to confidence measure. Medical students were slightly more certain than non-medical students of their estimations.

(3) There was a significant difference between medical and non-medical students on calibration in estimating the incidence of acute and prevalence of chronic disease conditions.

> Medical students were better calibrated than non-medical students in their estimations. Adjusting for age, students with medical training were not better calibrated than students without medical training. Holding the accuracy scores constant, the differences in calibration persisted.

Implications

The results of this study conflict with previous studies of calibration and confidence in three respects.

First of all, previous studies (Oskamp, 1962; Sanders, 1963; Hazard and Peterson, 1973; Lichtenstein and Fischhoff, 1976; Lichtenstein and Phillips, 1977) concluded that the most common bias in subjective probabilities is overconfidence. That is, people believe they know more than they actually know. In contrast, this study found that the problem with subjective probabilities was underconfidence.

Secondly, past studies (Lichtenstein and Fischhoff, 1976; Lichtenstein and Phillips, 1977) found that experts do not make better judgments and that they are overconfident in their estimations. The present study, however, found that experts (medical students) are not overconfident, but that non-experts (non-medical students are underconfident.

The final discrepancy between past studies and the present study is that previous studies (Lichtenstein and

Fischhoff, 1976; Lichtenstein and Phillips, 1977) found that expertise in a particular area does not lead to better calibration. This study, in contrast, found that experts are better calibrated than non-experts.

In order to explain the discrepancies between the findings in the literature and the results of this study, one must examine the differences in subjects used in previous studies, as compared with the present study.

First of all, the discrepancy between previous studies and the present study in the questions asked of the subjects. Previous studies (Oskamp, 1962; Lichtenstein and Fischhoff, 1976; Lichtenstein and Phillips, 1977) and the present study used the two-alternative approach, in which subjects were given two choices in responding to questions. They were asked to select the more likely alternative and to give the probability that their choices were correct. In stating the probability, Oskamp (1962), Lichenstein and Fischoff (1976) and Lichtenstein and Phillips (1977) used the half range method in which the response must be $\geq .5$ and the subject may use any response from .5 to 1. This was in contrast with this study which used the full-range method, in which the subject uses the range from 0 to 1 in determining the correctness of a response. The type of responses, whether half-range ($\geq .5$) or full-range approach (0 to 1.0), have considerable effect on the

results of the study. By limiting the subjects to responses between .5 and 1.0 for two alternative items, the subjects will be "forced" into the overconfidence side of the curve (See Figure 2). For example, when a subject chooses response A with a probability of only .2 of being correct, this means that he should have chosen response B with a probability of .8.

Secondly, a discrepancy lies in the type of subjects used. In attempting to determine if there was a difference between experts and non-experts in calibration and confidence, past studies (Oskamp, 1962; Sander, 1963; Lichtenstein and Fischoff, 1976; Lichtenstein and Phillips, 1977) used subjects who specialized in the same area but had different levels of expertise. For example, in Oskamp's study (1962) three different levels of expertise in psychology were used, namely, undergraduate psychology students, clinical psychology trainees and clinical psychologists. In another study by Sanders (1963) students of meteorology and instructors of meteorology were used. Again, all subjects had similar backgrounds, with differing levels of experience. Lichtenstein and Fischoff (1976) and Lichtenstein and Phillips (1977) used graduate students in a psychology department as subjects and divided them into three subgroups: best subjects, middle subjects and worst subjects, based on the accuracy of their responses in analyzing calibration

and confidence. The present study, however, used subjects with different specialities to compare experts to non-experts (that is, subjects with training in medicine versus subjects without training in medicine). Thus, the differences between groups were greater.

A third discrepancy was the nature of the task used. Past studies by Lichtenstein and Fischoff (1976) and Lichtenstein and Phillips (1977) used a sample of graduate psychology students to perform a task which was divided into two parts. One part represented expert knowledge (psychology items) and the other part represented non-expert knowledge (general knowledge items). Since no performance differences were found using these two parts, the study concluded that expertise did not contribute to calibration. The task of the present study involved only medically-related items; no general knowledge items were intended. Also, two distinct groups were used. Hence, any differences were due to difference in groups performing the same task.

Overall, then, the discrepancies in results obtained by this study compared with previous studies, are due to:

(1) the format of the questions asked;
(2) the types of subjects used; and
(3) the nature of the task performed.

Recommendations for Future Study

From the results and the discussion of the findings

the following recommendations for future study are suggested.

A replication of the present study with more than the two groups that were used (medical and non-medical students). It is recommended that practicing physicians be the additional group; since physicians have more experience in medicine it is assumed that the difference among the groups will be greater. Nursing students or graduate students in public health could be another group, as they should have considerable expertise in epidemiology. Also, it would be interested to match the age of medical and non-medical students when selecting the sample for analysis.

Another future study might be a replication of the present study using a revised task with more specific medical knowledge. For example, using biological science instead of epidemiological questions, a greater difference in medical knowledge between the two groups might be more clearly shown.

A longitudinal study could also be conducted following the medical students through their four years of training in the medical school to determine when a pattern of differences occurs in the accuracy and calibration of their decision-making. In other words, what are the trends in accuracy, confidence and calibration throughout the years of medical training? Are these trends linear, cubic or quadratic? A cross-sectional study can also be done if

there is evidence that indicates cohorts in different years are equivalent in terms of access to knowledge. In such case a cross-sectional study will be more appropriate and faster.

Another recommendation is to conduct interviews with the subjects of a replicated study to determine the factors on which they based their decisions in choosing responses to the items in the questionnaire. Some verbal (undocumented) comments from the non-medical students in this study were: "I don't know any of those answers." "Since my dad has diabetes, I tend to choose 'diabetes' rather than other conditions." Another subject in the non-medical group said, "I chose the answers to items based on what was reported on television or newspapers." A medical student in the pilot study stated, "I chose my answers based on the cases I encountered in the hospital during my internship training."

From the above comments the following question was raised: What distinct factors influence the medical versus non-medical students to choose one over another? If there is a distinct difference, does media reportage direct experience with the disease (that is, relatives or friends who have the disease) determine their choice of the more common disease condition among two given disease conditions? For medical students, is it training or experience that determines their choice of one answer rather than the other? Further studies that seek to answer the above questions can

give insight into the decision-making process and can be referred to as process-tracing studies.

This particular study concerned the medical field but its procedures are not inherently related to medical decision-making. It would be beneficial to conduct the same research in other specialized areas to see if the results are the same.

In summary, this study is a first experiment of a series of proceeding experiments. It is hoped that it will stimulate similar research studies to be conducted in the future so that more insight can be gained in determining the effect of training on decision-making.

BIBLIOGRAPHY

Beach, B.H. Expert judgment about uncertainty: Bayesian decision-making in realistic settings. Organizational Behavior & Human Performance, 14, 10-15 (1975).

Cooley, Donal G., ed., Better Home and Gardens Family Medical Guide. Meredith Cooperation, 1973.

Edwards, W. and Tversky, A. Decision Making. Baltimore, Penguin, 1967.

Ferrell, W.R. and McGoey, P.J. A model of calibration for suggestive probabilities. Organizational Behavior and Human Performance, 1975, 13, 1-16.

Fischhoff, B., and Beyth R. I knew it would happen - remembered probabilities of once - future things. Organizational Behavior and Human Performance, 1975, 13, 1-16.

Fischhoff, B.; Slovic, P. and Lichtenstein, S. Knowing with certainty: the appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance. Vol. 3, No. 4, pp. 552-564, 1977.

Gustafson, D.H. Comparison of methodologies for predicting and explaining hospital length of stay. Unpublished doctoral dissertation. University of Michigan, 1966.

Gustafson, J.E. The computer use in practice. Proceedings of the Fifth IBM Medical Symposium, 101-111, 1963.

Hazard, T.H. and Peterson, C.R. Odds versus probability estimates in a medical decision-making problem. (Michigan Mathematical Psychology Program, Report

Harris, Michael. Five counterrevolutionists in higher education. Corvallis: Oregon State University Press, 1970.

Kaheman, D. and Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237.251.

Lichtenstein, S. and Fischhoff, B. Do those who know more also know more about how much they know? Oregon Research Institute Research Bulletin, 16, (1), 1976.

Lichtenstein, S., Phillips, L. Calibration of probabilities: the state of the art. In H. Jungermann and G. de Zeeuw, (eds.). Decision making and change in human affairs. Dordrecht-Holland, Reidel Publishing Co., 1977.

Lichtenstein, S.; Slovic, P.; Fischhoff, B.; Layman - Combs, B. Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and memory, Vol. 4, No. 6, 651-678, 1978.

Lichtenstein, S.; Fischhoff, B. Training in calibration. Organizational Behavior and Human Performance, 26, 1980.

Murphy, A.H. and Winkler, R.L. Subjective probability forecasting experiments in meteorology: some preliminary results. Bulletin of the American Society, 55, 1206-1216, 1974.

Oskamp, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs, 75, (28, Whole No. 447), 1962.

Peterson, C.R. and Beach, L. R. Man as an intuitive statistician. Psychological Bulletin, 68, 29-46, 1976.

Pickhardt, R. C. and Wallace, J.B. A study of the performance of subjective probability assessors Decision Sciences, 5, 347-363, 1974.

Pratt, J.W. Personal communication in Lichtenstein, S. and B. Fischhoff's calibration of probabilities: the state of the art. In H. Jungermann and G. de Zeeuw (eds.). Decision-making and Changes in Human Affairs, Dordrecht-Holland: Reidel Publishing Company, 310-311, 1977.

Sanders, F. On subjective probability forecasting. Journal of Applied Meteorology, 2, 191-201, 1963.

Schaefer, R.E. and Borcherding, K. The assessment of subjective probability distribution: A training experiment. Acta Psychology, 1973, 37, 117-129.

Simon, H.A. Administrative Behavior. New York: Macmillan, 1945.

Stael Von Holstein, C.A.S. An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 10, 634-645, 1971.

Stael Von Holstein, C.A.S. Probabilistic forecasting: an experiment related to the stock market. Organizational Behavior and Human Performance, 8, 1939-158, 1972.

Tversky, A. and Kahneman, D. Judgment under uncertainty. Science, 185, 1124-1131, 1974.

Winkler, R. L. The qualificatin of judgment: some experimental results. Proceedings of the American Statistical Association, 386-395, 1967.

Winkler, R.L. Probabilistic prediction: some experimental results. Journal of the American Statistical Association, 66, 675-685, 1971.

APPENDIX A

FIFTEEN SELECTED PAIRS OF ACUTE CONDITIONS USED IN PART I OF THE QUESTIONNAIRE

The six acute conditions chosen are:

1. Disease of the ear
2. Fractures and Dislocations
3. Headache
4. Influenza
5. Pneumonia
6. Sprains and Strains

The 15 pairs of acute conditions are as follows:

1,2	2,3	3,4	4,5	5,6
1,3	2,4	3,5	4,6	
1,4	2,5	3,6		
1,5	2,6			
1,6				

APPENDIX B

FIFTEEN SELECTED PAIRS OF CHRONIC CONDITIONS USED IN PART I OF THE QUESTIONNAIRE

The six chronic conditions chosen are:

1. Arthritis
2. Cerebrovascular
3. Diabetes
4. Epilepsy
5. Heart conditions
6. Tuberculosis

The 15 pairs of acute conditions are as follows:

1.2	2.3	3.4	4.5	5.6
1.3	2.4	3.5	4.6	
1.4	2.5	3.6		
1.5	2.6			
1.6				

APPENDIX C
A Questionnaire on Calibration of Medical Probabilities

Thank you for your participation in this project. Please complete this page before filling out the questions.

Name:_____(optional)

Major:_____

Class Level:_____(graduate or undergraduate)

How long have you been in your program? ____1st yr. ____2nd yr.

____3rd yr. ____4th yr. _____Others.

What degree are you working at: ___ M.D., ___ D.O, ___Ph.D., ____M.A.,

___M.S., ___B.A., ___B.S., ___non-degree, _____Others.

What is your age?_____

How many years have you been in the United States? 1 year_____, 2 years_____,

3 years_____, 4 years_____, 5 years_____, 6 years_____, 7 years_____,

8 years_____, 9 years_____, 10 years_____, Over 10 years_____.

How many years of formal education have you had since high school? 1 year_____,

2 years_____, 3 years_____, 4 years_____, 5 years_____, 6 years_____,

7 years_____, 8 years_____, 9 years_____, 10 years_____, Over 10 years____

PART I - Acute Conditions: Incidence of New Cases

Each item consists of a pair of acute medical conditions. Acute condition is one which has been noticed less than 3 months (recent onset) that has either demanded medical attention or has restricted activities. The acute conditions included in this investigation are:

Influenza
Pneumonia
Headache
Fracture and Dislocation
Sprain and Strain
Disease of the ear

More information about these conditions is given in a separate sheet for your reference.

In each item, you will be given a pair of conditions. The question we want you to answer is:

Which one of the two conditions has more <u>new</u> cases occurring per year among people in the United States?

For each pair of acute conditions, we want you to <u>circle</u> the one that you think has more new cases occurring among people in the United States. For each answer, you might be very certain or uncertain that your response is correct. We want you to indicate how confident you are that your chosen answer is correct by circling the appropriate percentages of certainty. The percentages of certainty range from 0% to 100% with 0% as absolutely uncertain that your answer is correct and 100% as absolutely certain that your answer is correct.

For example:

A. Headache
B. Influenza
How confident are you that your answer is correct?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

If you think headache has more new cases occurring than influenza, you <u>circle</u> A. For the second part of the question, if you are absolutely sure that your answer is correct, you <u>circle</u> 100%. If you think that the chance of your answer being correct is only 10%, <u>circle</u> 10% or if you are absolutely uncertain if your chosen answer is correct, you can <u>circle</u> 0%. You can make your judgment on how confident you are that your answer is correct by circling any <u>one</u> of the percentages on the scale ranging from 0% to 100%.

INSTRUCTIONS:

1. Circle either A or B to indicate which one of the two conditions has more new cases occurring in the United States.
2. Circle the appropriate percentages to indicate how confident you are that your answer is correct.

1. A. Headache
   B. Sprains and strains

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

2. A. Influenza
   B. Pneumonia

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

3. A. Diseases of the ear
   B. Fractures and dislocations

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

4. A. Headache
   B. Pneumonia

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

5. A. Fractures and dislocations
   B. Pneumonia

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

6. A. Diseases of the ear
   B. Influenza

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

7. A. Fractures and dislocations
   B. Sprains and strains

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

8. A. Pneumonia
   B. Diseases of the ear

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

9. A. Influenza
   B. Fractures and dislocations

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

10. A. Sprains and strains
    B. Pneumonia

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

11. A. Headache
    B. Diseases of the ear

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

12. A. Sprains and strains
    B. Influenza

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

13. A. Headache
    B. Fractures and dislocations

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

14. A. Diseases of the ear
    B. Sprains and strains

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

15. A. Influenza
    B. Headache

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

16. A. Pneumonia
    B. Influenza

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

17. A. Fractures and dislocations
    B. Diseases of the ear

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

18. A. Pneumonia
    B. Sprains and strains

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

19. A. Fractures and dislocations
    B. Headache

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

20. A. Headache
    B. Influenza

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Definitions of Acute Conditions in Part I

Influenza:	An acute viral infection of the respiratory tract. Symptoms appear abruptly and include fever, chills, a dry cough, nasal stuffiness, running nose, aches and pain all over the body, a sore throat, headache, loss of appetite, nausea, weakness and depression.
Pneumonia:	Inflammation of the lungs from bacterial fungal or viral infection or from chemical damage. The outflow of fluid and cells from the inflamed lung tissue fills the airspaces, causing difficulty in breathing.
Headache:	Pain in the head.
Fractures and Dislocations:	Fracture is the breaking of bone(s). Dislocation is the displacement of bone(s) at a joint from its normal position.
Sprain and strain:	Sprain is the tearing of ligaments that hold bones together at a joint. Strain is an injury to a muscle or its tendon.
Disease of the ear:	Any disorder related to the ear or the auditory nerve and its central connections. Ringing in the ears and deafness are frequent symptoms suggesting disease of the ear.

PART II - Chronic Conditions: Prevalence of Existing Cases

In Part II, each item consists of a pair of chronic medical conditions. A chronic condition is defined as a condition which has been noticed more than 3 months that has either demanded medical attention or has restricted activities. The chronic conditions included in this investigation are:

Arthritis
Cerebrovascular disease
Diabetes
Epilepsy
Heart conditions
Tuberculosis

More information about these conditions is given in a separate sheet for your reference.

You will be given a pair of chronic conditions in each item, A and B. The question we want you to answer is:

Which one of the two conditions has more cases existing among people in the United States? That is, which condition is more prevalent among people in the United States?

For each pair, you might be very certain or uncertain that your response is correct. Please indicate how confident you are that your chosen answer is correct by circling the appropriate percentage of certainty. The scale for the degree of certainty ranges from 0% to 100% with 0% as absolutely uncertain if your answer is correct and 100% as absolutely certain that your answer is correct.

For example:

A. Diabetes
B. Epilepsy
How confident are you that your answer is correct?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

If you think that diabetes might have more cases existing than epilepsy in the United States, circle A. For the second part of the question, if you are absolutely sure that your answer is correct, you circle 100%. If you think that the chance of your answer being correct is only 10%, circle 10% or if you are absolutely uncertain if your answer is correct, you can circle 0%. You can make your judgment on how confident you are that your answer is correct by circling any one of the percentages on the scale ranging from 0% to 100%.

INSTRUCTIONS:

1. Circle either A or B to indicate which one of the two conditions have more cases existing in the United States.
2. Circle the appropriate percentages to indicate how confident you are that your answer is correct.

1. A. Arthritis
   B. Diabetes

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

2. A. Epilepsy
   B. Cerebrovascular disease

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

3. A. Arthritis
   B. Heart conditions

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

4. A. Heart conditions
   B. Tuberculosis

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

5. A. Diabetes
   B. Epilepsy

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

6. A. Cerebrovascular disease
   B. Tuberculosis

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

7. A. Diabetes
   B. Heart conditions

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

8. A. Epilepsy
   B. Arthritis

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

9. A. Diabetes
   B. Tuberculosis

   How confident are you that your answer is correct?

   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

10. A. Cerebrovascular disease
    B. Diabetes

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

11. A. Tuberculosis
    B. Epilepsy

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

12. A. Heart conditions
    B. Epilepsy

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

13. A. Arthritis
    B. Cerebrovascular disease

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

14. A. Heart conditions
    B. Cerebrovascular disease

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

15. A. Arthritis
    B. Tuberculosis

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

16. A. Cerebrovascular disease
    B. Arthritis

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

17. A. Diabetes
    B. Cerebrovascular disease

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

18. A. Epilepsy
    B. Diabetes

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

19. A. Epilepsy
    B. Heart conditions

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

20. A. Tuberculosis
    B. Heart conditions

    How confident are you that your answer is correct?

    0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Definitions of Chronic Conditions on Part II

Arthritis:	Joints may be affected by inflammatory or degenerative changes which cause pain and stiffness described as arthritis.
Cerebrovascular disease:	Any disorder of the circulation of blood in the brain. The most common is disease of the blood vessel wall or clotting of the blood within the vessel.
Diabetes:	Diabetes is an inherited disease which occurs when the body cannot make full use of some of the foods we eat, mainly the carbohydrates or sugars and starches. The pancreas, a large gland lying beneath the stomach, does not make available enough insulin to burn these foods as energy or store them for future use. Starches and sugars increase the blood sugar content until the sugar passes through the kidneys and into the urine. The loss of carbohydrate energy causes enormous thirst and compensating urinary outflow.
Epilepsy:	A nervous disorder of varying severity, marked by recurring explosive discharge of electrical activity of brain cells producing convulsions, loss of consciousness, or brief clouding of consciousness.
Heart conditions:	Disorders of heart affect its ability to pump blood around the body which cause deep aching, crushing or vice like pain in the chest, radiating perhaps the arm or the neck or jaw. Active rheumatic fever, chronic rheumatic heart disease, hypertensive heart disease, coronary heart disease are all heart conditions.
Tuberculosis:	An infection caused by the bacterium mycobacterium tuberculosis. The most common site of infection is the lung. Symptoms of tuberculosis include malaise (a general feeling of being unwell), lassitude, tiredness, loss of appetite, fever and night sweats.

APPENDIX D

AN EXAMPLE OF HOW ACCURACY, OVER/UNDERCONFIDENCE AND CALIBRATION SCORES WERE DEVELOPED

Formulas

$$\text{Accuracy} = \frac{n}{N}$$

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

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

Key to Symbols

n = Number of items identified correctly

N = Total number of items

r_t = Assigned probability response

n_t = The number of times the response r_t was used

c_t = The proportion correct for all times assigned probability r_t

T = Total numbers of different response categories used

An Example

r_t	Correct/Incorrect Tally (1=cor, 0=incor.)	n_t	c_t	$r_t - c_t$	$n_t(r_t - c_t)$	$(r_t - c_t)^2$	$n(r_t - c_t)^2$
.0							
.1							
.2							
.3							
.4							
.5	011	3	2/3= .67	.50-.67= -.17	3(-.17) =-.51	$(-.17)^2$= .028	3(.028) =.084
.6	11100	5	3/5= .60	.60-.60= 0	5(0) = 0	$(0)^2$= 0	
.7	11	2	2/2=1.00	.70-1.00=-.30	2(-.3) =-.60	$(-.3)^2$ = .09	2(.09) =.18
.8	01110111	8	6/8= .75	.80-.75 = .05	8(.05) = .40	$(.05)^2$ =.0025	8(.0025)=.02
.9	1101	4	3/4= .75	.90-.75 = .15	4(.15) = .60	$(.15)^2$ =.0225	4(.0225)=.09
1.0	11111111	8	8/8=1.00	1.00-1.00= 0	8(0) = 0	$(0)^2$ = 0	
					$\sum_{t=1}^{T} n_t(r_t - c_t)$=-.11		$\sum_{t=1}^{T} n_t(r_t - c_t)^2$=.374

$$\text{Accuracy} = \frac{n}{N} = \frac{24}{30} = 80$$

$$\text{Over/Underconfidence} = 1/N \sum_{t=1}^{T} n_t(r_t - c_t) = \frac{-.11}{30} = -.0033$$

$$\text{Calibration} - 1/N \sum_{t=1}^{T} n_t(r_t - c_t)^2 = \frac{.374}{30} = .01246$$