6-1-1969

Physicians' Degree of Concern as a Function of Numerical Medical Information Indices

Frederick W. Theye

PHYSICIANS' DEGREE OF CONCERN AS A FUNCTION OF NUMERICAL MEDICAL INFORMATION INDICES

by

Frederick W. Theye

B. S. Concordia Teachers College, 1963

M. A. University of North Dakota, 1965

A Dissertation

Submitted to the Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the Degree of

Doctor of Philosophy

Grand Forks, North Dakota

June
1969

This Dissertation submitted by Frederick W. Theye in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota is hereby approved by the Faculty Advisory Committee under whom the work has been done.

L. R. O. [illegible]

R. H. Kolstoe

E. S. Halas

John O. Noll

Wm. I. Morgan

A. William Johnson
Dean of the Graduate School

Permission

Title Physicians' Degree of Concern as a Function of Numerical Medical Information Indices

Department Psychology

Degree Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairman of the Department or the Dean of the College in which my dissertation work was done. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Signature ____________________

Date ____________________

ACKNOWLEDGEMENTS

This writer wishes to express his deepest appreciation to Dr. LeRoy Stone whose original idea provided the basis for this investigation. His generous contribution of time, effort, and guidance certainly facilitated the completion of this inquiry.

Appreciation is also extended to Drs. Ralph Kolstoe, Edward Halas, John Noll, and William Morgan for who also freely gave of their time and experience. The writer also wishes to thank Drs. Barcome and Harwood who provided many helpful suggestions in the early formation of this study.

Finally, a special expression of gratitude is extended to my wife Darlyce, and sons David and Darren, all of whom made numerous sacrificies throughout my graduate program. Their constant support, comfort, and assurances provided a strong motive for completing my graduate education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ABSTRACT

The present investigation was essentially exploratory and descriptive in purpose. Utilizing the direct scaling method of magnitude estimation (with assigned modulus) the study attempted to: (1) determine if lawful relationships existed between the clinical judgment of licensed physicians and the results from nine numerical medical information indices, and (2) to quantitatively describe such relationships.

It was believed that direct estimation methodologies have been shown through empirical studies to be superior to the psychophysical models of Fechner and Thurstone. Previous work has also illustrated that the power law of S. S. Stevens' has provided a powerful methodology in studying the topic of clinical judgment.

In the present study 27 licensed physicians served as judges. They judged results from nine frequently used numerical medical information indices which were varied systematically and independently. An upper and lower limit for each of the nine indices was determined from the medical literature and medical consultants. Specific stimuli within these limits were spaced in equal logarithmic steps when feasible. Judgments were made relative to degree of concern for a contrived 35 year old patient's health status. The laboratory test-indices and the various levels of each test were presented in randomized orders.

For four of the indices, levels above as well as below normal were included. In scale development, these were considered separately, thus 13 subjective scales were developed.

In general, the results indicate that for nine of these scales the relationships observed were curvilinear when degree of concern was plotted against the appropriate stimulus metric. A log-log transformation rectified the data so that straight lines offered reasonably good approximations of the observed trends. It was determined that a power function model was an appropriate description of these data. For four indices the relationships, when degree of concern was plotted against the stimulus continuum, were markedly linear in nature. It was suggested that: (1) the underlying continua for these four indices may be metathetic, or (2) that physicians view these four indices as some sort of ordered category measures even though the underlying stimulus measures are continuous in nature.

Implications for the direct estimation literature seem clear. This study represents one of the earliest successful extensions of these measurement methodologies into the topic of clinical judgment. It was suggested that direct estimation procedures are sufficiently sensitive to assist in the clarification of the many enigmatic ambiguities now existant in the clinical judgment literature.

Implications for medical education were also drawn. The development of scales similar to those produced in this inquiry could provide valuable communication vehicles whereby the "exigencies of the office would be brought into the classroom."

For several scales the predetermined standard was believed to be disparet from the intrinsic standard employed by the judges. This was believed to increase variability or noise in the measurement system. Inter-scale comparisons were also made and four of the indices seemed more potent in terms of eliciting concern. One index appeared to elicit relatively little concern. Data derived in connection with inter-scale comparisons holds potential for future research into this area.

Limitations of the present inquiry were discussed. For example, the sample was in no way random or systematic, and several standards which were employed seemed to be inappropriate. Suggestions for future research were also advanced.

CHAPTER I

INTRODUCTION AND STATEMENT OF THE PROBLEM

"The methods of psychophysics are ordinarily designed to solve problems related to the nature of organisms. The focus of interest is typically the normal observer, his thresholds, his resolving powers, and the magnitudes of his perceptions"(Stevens, 1958, p. 193). The psychophysical method, in general, is concerned with the study of stimulus-response relationships. Historically, the method can be traced to the pioneering work of G. T. Fechner in the mid-nineteenth century who, following the earlier work of E. H. Weber, carefully developed and defended his approach to the study of discriminal processes. Traditionally, the method has been used to study questions like, "how do organisms discriminate differences in the physical world," "the presence or absence of a stimulus," "the sensitivity limitations of organisms," etc. Fechnerian psychophysics was largely confined to what many felt were matters of little consequence, and, in fact, "psychophysics has been viewed intrinsically as one of the more 'ivory tower' areas of experimental psychology" (Stone, 1968a, p. 161). The importance, however, of this model of measurement should not be underestimated since it has had a most pervasive effect, and it has "set experimental quantitative psychology off upon the course which it has followed" (Boring, 1957, p. 294).

Historical Approaches in Psychophysics

Fechner's law, based on a logarithmic model, asserts that sensation increases in arithmetic steps as the stimulus magnitude increases in ratio steps. This statement represented the first psychophysical model, and it had a significant impact despite the fact that considerable controversy swirled around it. It remained the only visible law for some 67 years until Thurstone proposed his law of comparative judgment in 1927. This law was an extension and elaboration of the Fechnerian model, however, it incorporated some contemporary psychometric concepts into the law. Thurstone's law was widely utilized as it successfully demonstrated its value in studying the more "applied" areas of attitudes and opinions. Since the model was an extension of a basic technique into areas that were of interest to wider audiences, it represented an important contribution. Until this time psychophysics had been used primarily to relate scale values of responses on a psychological continuum to stimuli on a physical continuum, i.e., stimuli were metric in nature. Thurstone eliminated the need for metric stimulus values, and his law was utilized in endeavors to scale psychological variables without the need for an underlying physical dimension.

Observed variability of human judgments provided the cornerstone of this model and its purpose, in general, was to move from units of variability to subjective magnitude units by way of various psychometric assumptions. Thurstone's hope was to establish a scale of equal intervals with an arbitrary zero-point like the ordinary scale of temperature (Stevens, 1966).

From the 1920's until the present time, attitudes, values, preferences, and the subjective impressions of subjects have been given increasing research attention. Many nonmetric continua which have no underlying physical dimensions have been successfully scaled following Thurstone's model.

The third major development in psychophysics occurred with S. S. Stevens' (1957) power function model which proposed that equal stimulus ratios produce equal subjective ratios. On numerous perceptual continua, direct assessments of subjective magnitude seem to bear an orderly relation to the magnitudes of the stimuli. To a fair first order approximation, the ratio scales constructed by direct estimation methods are related to the stimuli by power functions of one degree or another. The idea that equal stimulus ratios produce equal subjective ratios has been entitled the power law of psychophysics, and it stands at variance with the Fechnerian and Thurstonian laws in several important respects.

The first such difference lies in the Fechnerian and Thurstonian proposition that scale units can be developed from observations of variability. Stevens (1959, p. 389) notes that by "processing data on confusions, just noticeable differences (jnd's), average errors, . . . the members of this school propose to erect interval scales of psychological magnitude." That is, they attempt to "unitize dispersion." The direct estimation methods avoid this proposition, and assume only that the observer is capable of following the instructions to make ratio judgments.

A second major difference between the power law position and those of Fechner and Thurstone centers on whether the approach to the observer is direct or indirect. Ekman and Sjöberg (1965), in differentiating between the direct and indirect methods, note that in the Thurstonian method:

> . . . only a minimum of information is required and obtained from the subject -- essentially rank order. Because of the lack of metric information, the scale is obtained from the experimental data by means of a set of assumptions, and thus the scale may be considered 'indirect.' The assumptions are concerned with variability -- over trials for a given subject or usually over subjects for a given trial. (p. 451)

In contrast, the direct methods operate on the basic assumption that the subject operates in accordance with the instructions, and the scale construction is a straightforward procedure essentially consisting of averaging experimental data.

The power law concept has, for many scientists, also forced a revision of Fechner's assumption that jnd's are subjectively equal; and the Thurstonian assumption, which parallels Fechner's assumption, that discriminal dispersions are constant up and down the scale. By way of contrast, Stevens (1959, p. 389) posits that "discriminal dispersions grow directly in proportion to the psychological magnitude."

The new direct estimation methods have revolutionized psychophysical research, and their introduction has revitalized research with scaling methods. In fact, "99% of all work dealing with problems of scaling, or the application of scaling methods to psychological problems, has been published since 1950 . . . . " (Ekman and Sjoberg, 1965, p. 451.)

Scalable Continua

Stevens (1959) considers two types of continua which are amenable to scaling procedures: prothetic and metathetic. Traditionally, dichotomous distinctions are made between quality and quantity or size versus sort. Distinctions like these are similar to the prothetic-metathetic distinction offered by Stevens, and are supported by convincing empirical evidence from other independent investigators (e.g., Perole, 1963; Eisler, 1962). Prothetic continua are concerned with the general quantitative questions such as "how much," whereas metathetic continua have to do with the qualitative questions of "what kind and where." Discriminations on some prothetic sensory continua appear mediated by an additive process at the physiological level as seen in loudness, heaviness, brightness, etc.; where progress along the continuum is accomplished by adding excitation to excitation. In contrast, discriminations along metathetic sensory continua appear substitutive in nature as seen in the phenomena of pitch, position, etc.; where progression along the continuum is achieved by changing the site of stimulation (Stevens, 1957). Determination of these suggested differences at a physiological level is most difficult at this time, and Stevens suggests four other more "functional" criteria to differentiate between prothetic and metathetic continua.

These functional criteria are: (1) the subjective size of the jnd's; (2) the shape of the relationships between scales obtained by the direct and indirect methods; (3) time-order errors; and (4) the hysteresis phenomenon. Jnd's are not equal in subjective size along prothetic continua as they are along metathetic continua. For example, the sensation produced by a stimulus 50 jnd's above threshold is not half as great as one produced by a stimulus 100 jnd's above threshold which would be the implication in Fechnerian psychophysics. Rather, according to Stevens (1957, p. 154), "the hard fact is that if the typical subject were confronted with two such stimuli on a Class I (prothetic) continuum he would assert with certainty that the ratio between the two sensations is greater than two, because scales obtained by summating jnd's are nonlinearly related to scales of subjective magnitude."

The second functional criterion deals with the relational shapes of scales obtained by direct and indirect methods. Category rating scales or partition scales are functions obtained when subjects judge sets of stimuli with respect to categories identified by numbers or adjectives. Although category rating scales are based on a direct form of measurement, they require a judge to partition the subjective scale into equal units. Stevens and Galanter (1957), reporting results based on 12 perceptual dimensions note that, for prothetic continua, category scales (as the ordinate) are concave downward when plotted against a ratio scale of subjective magnitude. Metathetic continua may be linear when so plotted. The reason for this phenomenon appears related to the subjects' inability to equalize intervals in their category scales due to variation in their sensitivity. That is, they are less sensitive at higher stimulus levels; therefore, they are not able to equalize the intervals even when so instructed.

The time-error constitutes the third functional criterion, and it refers to the fact that the second of two equal stimuli is usually judged to be greater than the first. Stevens (1957) notes that they have reason to believe that a systematic time-error is typically characteristic of prothetic and not metathetic continua. Although "it is important to note that the error on prothetic continua is typically small - a fraction of a jnd . . . " (p. 157). The difference observed between the two continua again is believed related to the phenomenon of sensitivity asymetry discussed in the preceding paragraph.

The fourth functional criterion utilized in distinguishing between the two kinds of continua is hysteresis which means a "lagging behind." Stevens (1957) notes that it seems to be a good term to describe what happens when the apparent sense distance between successive stimuli is judged in different orders. For example, it is as if the loudness the subject hears lags behind what he should hear as he goes up and down the scale. The experimental results surrounding this criterion are more equivocal than with the other three, and they can be considered only suggestive... But they would seem to indicate that the hysteresis phenomenon occurs on prothetic and not on metathetic continua.

In summary, the four criteria discussed above have been shown through empirical studies (Stevens, 1957; Stevens, 1966; Stevens and Galanter, 1957) to be of varying value in making the distinction between metathetic and prothetic continua. While several researchers (e.g., Warren and Warren, 1963, and Torgerson, 1960) have voiced some questions about the validity of these functional criteria, considerably more empirical evidence will be needed before definitive conclusions can be drawn.

Advantages of Direct Estimation Ratio Scaling

Aside from the theoretical considerations discussed above, the direct estimation ratio scaling methods offer a number of substantial advantages to the researcher when compared to the indirect scaling models of Fechner and Thurstonian partition model. The advantages to be discussed are four: (1) a higher level of measurement can be obtained; (2) the reliability of the measure is high; (3) contextual effects are more easily controlled; and (4) the "new" psychophysical methodologies and concepts are easily applicable in nonmetric stimulus situations (Stevens, 1966a).

The various ratio scaling techniques such as magnitude estimation produce scales at the ratio level of measurement. This is the highest level of numerical measurement (cf. Stevens, 1951), and with such it is possible to carry out any arithmetical operation (transformation) desired. By way of contrast, category scales result in measurements of essentially rank order (ordinal level) which restricts the type of arithmetical operations that can be conducted with such numerical data. A ratio level of measurement permits one to state, for example, not only that B possesses more of a given characteristic than A (rank order); but also allows that B has three times as much of the specific characteristic as A. This latter is considerably more potent a statement in terms of the amount of information communicated, and potentially is much more useful in attempting to understand a numerically measured phenomenon.

Secondly, direct estimation methods have provided a highly reliable index of stimulus-response relationships. This has been demonstrated in numerous laboratories employing a variety of stimuli and response categories. For example, Stevens and Galanter (1957) asked judges to make subjective judgments concerning brightness of light viewed in a dark room; Ekman and Künnapas (1963) successfully constructed scales pertaining to political importance of Swedish monarchs; Koh (1965) developed a ratio scale using psychiatric patients as judges and asked them to make esthetic judgments of music; Hermann and Fox (1967) successfully employed magnitude estimation procedures to scale attitudes regarding sexual standards; and Stone (1968) scaled psychiatric judgment relative to severity of impairment of functional psychotic disorder classifications.

Thirdly, context effects can be easily minimized with direct estimation scaling methods, and they are extraneous variables which contaminate experimental results. These effects are omnipresent in many laboratory and clinical situations. Stevens (1966) presents not only the results of his studies but those of independent laboratories (e.g., Fillenbaun, 1963) to support his contention that direct estimation methods are not overly sensitive to context effects.

Fourthly, numerous nonmetric stimuli have been scaled using the "new" psychophysical methodologies. Indow (1959) presented Japanese university students with pictures and descriptions of pairs of watches. They were to state a preference and then to indicate the relative preference strength in ratio judgments. Sellin and Wolfgang (1964) successfully related specific types of delinquent behaviors and preceived subjective seriousness of the offenses through ratio scaling methodologies. More recently, Stone (1968b) successfully scaled psychiatric judgment relative to the degree of constitutionality in various functional psychoses. In general, a host of other investigations have indicated that stimuli which have no discernible underlying metric can be successfully scaled employing direct estimation procedures.

The empirically demonstrated utility of the direct estimation techniques allows one to utilize the wealth of knowledge derived from classical and modern psychophysics in the clinical setting (Stone, 1968b). This is no small advantage when the vast empirical history of the classical psychophysical method is considered. As Hunt (1962, p. 48) notes, "If our basic (judgment) processes are indeed similar to those of psychophysics, we can profit from an extensive literature on scale construction."

Investigation of Clinical Judgment

The topic of clinical judgment or decision making is lively and controversial one at the present time (Goldberg, 1968), although surprisingly little research effort has been conducted in the area relative to its central role.

For the physician, clinical judgment typically involves three separate yet interdependent information sources: physical examination, routine laboratory test, and the clinical history (Sodeman, 1964). Each of these sources can provide valuable data relevant to the goal of approaching the patient in a therapeutic manner. The process employed by physicians in extracting relevant information from such sources has been actively studied through a variety of techniques, and from a myriad of approaches. Ledley and Lustead (1959) were among the earlier advocates of analyzing physicians' judgment processes into separate parts and relating these processes to computer functioning. The interest in computerizing medical information has grown rapidly, as seen in the Kaiser Foundation program and certainly will continue to receive attention (Lipkin, 1964, Erdman, 1964). Collen (1967, p. 4) predicts that "In the future . . . it is likely that larger hospitals in every community of 100,000 or more will be affiliated with an automated multitest laboratory."

Rimoldi (1964) and his associates have conducted a sequence of studies in which they developed a series of pencil and paper tests to appraise medical diagnostic skills of physicians at various level of training. The tests employed both real clinical cases and contrived cases about which the subjects were to ask questions en route to a final diagnosis. They found that the number of questions asked decreases progressively from junior through seniors to practicing physicians, and that, although the juniors asked the most questions, they gain less information than do practicing physicians who asked the least number of questions. With some success, Adams (1964), analyzed tape recorded diagnostic teaching sessions so as to study the problem solving approaches of the instructor, the student, and their interaction, although considerable data had yet to be analyzed.

The computer analogy approach, pencil and paper testing, and teaching of problem solving strategies have all been employed by researchers in attempts to clarify, to understand, and to potentially improve the clinical judgment process of physicians. Block (1964), commenting on research strategies in this area of clinical judgment, suggests the presence of two divergent views: "one group . . . is inclined to relegate clinical decision making to the realm of the artistic . . . while the opposing group views clinical judgments as rational and scientifically verifiable." (p. 172) It is the suggestion of this paper that methodologies such as direct scaling techniques will support and buttress the claims of the latter camp and will prove profitable in resolving many questions pertaining to clinical judgments.

Clinical Psychophysics

As noted earlier, psychophysics traditionally has been associated more with basic research than with applied concerns. Until recently its methods have not been widely applied to clinical settings, although historically, test designers such as Binet employed psychophysical thinking when developing psychometric measurement devices. These methods have also been utilized in the narrow areas of clinical audiometric and visual testing (Stone, 1968a).

For many years there has been the tendency to view the clinical judgment process as a "special means of knowledge" or an intuitive procedure not readily amenable to empirical scrutiny. This sentiment, however, is being quickly dispelled as more and more research efforts into the area bear fruit. The analogy drawn between psychophysical and clinical judgment has been a valuable one, and the "investigation of categories of report in the field of clinical judgment is a lively, exciting area which promises much for the future" (Hunt, 1962, p. 49). Some authors (Meehl, 1954) have suggested that clinical judgment best be left to actuarial methods. However, Stevens (1958, p. 194), cognizant of the historical divorce between psychophysics and judgment processes, optimistically notes that, "Despite the ingenuity of modern instrumentation, many tasks of rating, grading, and judging can still best be done by two-legged meters . . . little of this type of activity gets attention in the academic laboratory, although much could probably be learned from its systematic study." Although made in the context of sensory psychophysics, Stevens' thought seems relevant to the area of clinical psychophysics. Certainly the challenge to explore the judgment process further has been set, and the embryonic beginnings of these explorations can be found in the recent literature.

The earliest concerted, systematic application of psychophysical procedures in areas such as clinical judgment was made by Hunt (1959) and his associates as early as World War II (Hunt and Jones, 1962). Hunt and Jones (1962) believe that there is a close relationship between clinical and psychophysical kinds of judgment. They suggest:

> They (clinical and psychophysical) are merely the opposite poles of a rough continuum, a quantitative continuum marked by the clarity of specificity with which the stimuli are designed, by degree to which the judgmental setting is standardized through careful control of the known pertinent variables and the elimination of extraneous cues, and by the provision of uniform modes of reporting . . . (p. 34).

While the efforts of Hunt and his associates are significant in terms of their application of psychophysical methods to clinical material they contain all the limitations inherent in partition scales.

In contrast with ratio scaling methodology, category scales make the assumption that variability remains constant regardless of stimulus magnitude. This has not withstood the empirical test (Stevens, 1966). Category scales also result in a lower order level of measurement. Finally, they typically result in a lower level of reliability as opposed to ratio scales' higher level of judgmental reliability.

Stone (1968a) was the first to utilize the term, clinical psychophysics. He notes:

> It would not seem to represent a travesty upon the name of psychophysics to speak of a clinical psychophysics. It does seem that the theory and methods of psychophysics, especially the newer direct estimation methods associated with the psychophysical power law, can be constructively utilized to better explore the judgmental continua involving clinical content (p. 172).

On this premise he conducted a series of studies (1966, 1968a, 1968b, 1969; Stone and Skurdal, 1968), employing direct estimation methods, concerned with psychiatric judgment of prognosis, constitutionality, predisposition, and degree of impairment for the 15 functional psychotic disorder classifications. In one of these studies, Stone and Skurdal (1968) found that a previously developed pair-comparison scale of prognostic favorability (Stone, 1966) was very "close to being a logarithmic function of the scale based on direct magnitude estimations" (p. 470). Stone (1969) then related this prognostic scale to three validity indices derived from research literature. The results illustrated that power functions approximately describe the relationships between judged prognosis: (1) the average improvement rates for these classifications; (2) median length of stay in the hospital; and (3) the median admission age.

This series of investigations has empirically demonstrated the promising utility of the "new" psychophysics when studying the clinical judgment process. With these studies having provided an empirically sound and demonstrable utilization of direct estimation techniques in the study of clinical judgment, the present investigation extended the use of these methods into the area of clinical-medical judgment.

Statement of the Problem

The present inquiry was essentially exploratory and descriptive in purpose. Utilizing the direct estimation method of magnitude estimation (with as assigned modulus) it attempted: (1) to determine whether lawful relationships exist between the licensed physicians' judgments of subjective concern and the results from nine rather routine laboratory tests (numerical medical indices), and (2) to qualitatively describe such relationships. Their judgments of subjective concern were scaled, and the relationships were examined for the extent, form and possible theoretical implications.

CHAPTER II

METHODOLOGY

Subjects

The judges (Js) were twenty-seven licensed physicians (all possessed the medical doctorate) from the states of Indiana (five), Minnesota (six), and North Dakota (sixteen). They were contacted on an individual basis by the investigator and asked if they would participate. The selection of Js was in no way systematic or random. The physicians from Indiana all practiced in the same medical clinic, and the Minnesota physicians were all on the staff of one hospital. Ten of the North Dakota physicians were employed at the State Hospital in Jamestown, and six were associated with the University of North Dakota School of Medicine.

The mean number of years in practice for the sample was 21.60 years with the range being from 1 to 47 years. The mean chronological age was 52.61, the range being from 33 to 73 years. The specialities and the number of Js in each were respectively: Pathology (five), General Practice (six), Psychiatry (eight), Obstetrics and Gynecology (two), Pediatrics (one), and Internal Medicine (five). Twenty-three of the Js were involved daily with the practice of clinical medicine, and the other four Js'chief responsibilities were in the area of medical education.

Stimuli

Nine different clinical laboratory tests or numerical medical indices were selected[1] on the basis that they represented frequently administered diagnostic clinical laboratory measures or indices requested by physicians for routine screening purposes. The tests were: white blood count (WBC), red blood count (RBC), temperature, pulse, systolic blood pressure, diastolic blood pressure, protein, sugar, and specific gravity of urine. Protein and sugar values were derived from urinalysis. The clinic-laboratory test-indices and the various levels of each test were presented in randomized orders to the Js.

For each clinical-laboratory test-index a range of possible values was determined by reviewing the relevant medical literature and by obtaining the opinions of medical consultants[2]. The consultants, in several instances noted that levels obtained from the literature were not "very pathologic." They suggested other "more pathologic" levels be utilized as upper and lower extremes of the stimuli ranges. When feasible, the specific values within the range were spaced in equal logarithmic steps for all tests except sugar and

---

[1]The indices selected were suggested by the medical consultants.

[2]The consultants were Dr. Donald F. Barcome, Professional Director of the Medical Rehabilitation Hospital, and Dr. T. H. Harwood, Dean of the University of North Dakota School of Medicine.

protein. These two tests are typically reported to physicians in terms of a six-step scale.[3]

The standards for each test-index were determined by selecting a value which was believed to be somewhat deviant from normal. The level which was one logarithmic step above normal limits was selected as the standard. This was done because the Js were instructed to make their judgments of concern relative to the standard. It was necessary therefore that a mildly "pathologic" level be presented as the standard since this was to represent or be associated with some degree of concern. The selection of the number 50 as the numerical modulus was intended to allow the Js a wide range of choices on either side of the modulus. Thus, they would be free to choose numbers larger or smaller than 50 to represent either greater or lesser degrees of concern than that represented by the standard (cf. Poulton, 1968).

The range and levels for each numerical medical information index were:

Red Blood Count: For adult males the normal range is considered to 5-6 million cu. mm. with an error rate of ±20% (Miller, 1955). The range utilized was from 1.0 to 13.0 million per cu. mm. The specific levels utilized were: 1.0, 2.5, 4.0, 6.0, 8.5, 10.0, 13.0. The standard was set as 7.0.

White Blood Count: For adult males it is generally agreed that a count of 5,000 to 10,000 is within normal limits considering an error rate of ±10% (Miller, 1955). The range utilized was from 750 per cu. mm. to 200,000 per cu. mm. The levels presented were: 750, 2,350, 5,250, 7,500, 9,500, 24,750, 100,000, 200,000.

---

[3]Test results are usually reported to physicians in terms of 0, Trace, 1+, 2+, 3+, 4+. This would seem to be a rough category scale although it is based on continuous data. For Sugar: Trace = 30 mg.%; 1+ = 30-99 mg.%; 2+ = 100-299 mg.%; 3+ = 300-999 mg.%; 4+ = 1,000 mg.%. For computational purposes the "trace" category was given a value of .30.

Note the later two values were not the result of logarithmic spacing but were included at the suggestion of the consultants. The standard was set at 10,500.

Temperature: The average adult temperature is 98.6° with a standard deviation of 0.50 (Sodeman and Sodeman, 1967). The range utilized in this study was from 97.0° to 108.2° with the levels in between logarithmically determined. The standard was 101.2° with the levels presented being: 97, 99.4, 102.6, 104.0, 105.4, 108.2.

Pulse: The average pulse rate for an adult male between ages of 30-35 is 70 (Altman and Ditmer, 1964). The range utilized in this study was from 40 to 200 beats per minute with the intermediate levels being determined logarithmically. The standard was set at 82, and the levels presented for judgment were: 40, 57, 68, 98, 118, 141, 168, 201.

Diastolic Blood Pressure: In healthy middle-age adults the average diastolic pressure is 80 mm (Harder and Gow, 1953). The range utilized in the present study was from 60 mm. to 226 mm. with intermediate levels being determined logarithmically. The standard was set at 93, and the levels presented were: 80, 108, 124, 145, 168, 195, 226.

Systolic Blood Pressure: In healthy middle-age adults the average systolic pressure is 120 mm (Harder and Gow, 1953). The range utilized in the present study was from 108 to 281 with intermediate levels being determined logarithmically. The standard was set at 130, and the levels presented were: 108, 130, 143, 173, 191, 211, 232, 252, 281.

Specific Gravity: The normal range for specific gravity of urine is 1.016 to 1.022. The range utilized was from 1.000 to 1.0400 with intermediate levels determined logarithmically. The standard was set at 1.0135, and the levels presented for judging were: 1.0000, 1.0036, 1.0102, 1.0168, 1.0201, 1.0267, 1.0300, 1.0400. The two extreme values were suggested by the consultants.

Sugar: Sugar content from urine is typically reported to physicians in terms of 0, trace, 1+, 2+, 3+, 4+. The standard was set at 1+, and the remaining five levels were presented to the J's.

Protein: Protein content in urine is typically reported to physicians in terms of 0, trace, 1+, 2+, 3+, 4+. The standard was set at 1+, and the remaining five levels were presented to the J's.

In discussions with the medical consultants it was clear that levels on the numerical medical indices should be related to a specified "patient." This was necessary because what is pathologic for one individual may not be so for another. Whether or not a particular level of an index is pathologic is partially dependent upon such factors as age, previous medical history, etc. A "patient" and his medical history were contrived to ensure that all Js would make their judgments of concern relative to the same patient. The Js were provided with the following clinical history:

> The patient is a 35 year old Caucasian male who comes to you in the morning for his annual physical examination. You have noted that he is alert, responsive, has good color, and is in good spirits. In addition, his gait is normal as is his posture, and he appears to be of average height and weight. He reports an essentially non-remarkable medical history, and volunteers the facts that he has never had major surgery, abnormal bleeding, nor any significant weight losses or appetite disturbances.

Instructions

The Js were presented with the following instructions enclosed within a manila folder:

> We would appreciate your cooperation in an experiment which will take only 10-15 minutes of your time. This is NOT an experiment designed to assess the accuracy or correctness of physicians' judgments. Rather, it is an attempt to quantify your expert clinical judgment of various laboratory test results. In deciding which lab tests to use we consulted with the Chief of Medical Services at the University of North Dakota Rehabilitation Hospital, and with the Dean of the Medical School, also at the University of North Dakota. We and they are cognizant of the fact that some of the situations presented in this study may be unusual in the sense that you typically would consider the results in relationship to some other data. However, we ask that you suspend this process for the study, and make judgments based solely on the data presented on the slips of paper.

Your judgments are to reflect the degree of concern you, as a physician, would have for the well-being of the patient described in the clinical history on the following page.

We ask that you make your judgments proportional to a standard which was set to assist you in making the judgments. For example, below you will find three lines of different lengths. Note that the standard has been set at 50. Now compare the length of line A with the standard, and judge its length proportional to the standard. For example, if you think it's 3 times as long you should put the number 150 (3 x 50) in the space provided. Now for line B also judge its length proportional to the standard.

STANDARD ______________________ 50

A ____ _____

B _________________ _____

Your figures were probably close to 100 for line A, and 35-40 for line B since the standard is 2 inches long, A is 4 inches long, and B is 1½ inches long.

The clinical history was presented here.

In the following pages you will note the results of various laboratory tests which are frequently given to many patients as part of a routine physical examination. Your task is to judge the degree of concern you, as his physician, have for this 35 year old man when the various lab results are changed in a non-systematic order. Assume that the results are based on the standard lab tests, and the analyses are correct! To assist you in making your judgments the degree of concern has been arbitrarily set at 50 when the lab results are as presented. Please assign numbers on the attached sheets in such a way as to reflect your degree of concern relative to the standard. For example, if you are twice as concerned when his temperature is reported as 104.5° as opposed to when his temperature is 100.2° (the standard being set at 50) you would put the number 100 in the space provided. If you are only one-fifth as concerned under these circumstances, you would place the number 10 in the space provided. For the remaining situations you may use <u>any numbers</u> you wish just be sure to make each judgment of concern <u>PROPORTIONAL</u> to the standard represented by the number 50. Please make each judgment independent of previous ones by simply turning each slip over after you have made your judgment.

Each medical index was presented on a single piece of 8½" x 11" paper. Attached to this single sheet were the "results" or levels of each test-index to be judged. Each level was on an independent slip of paper. The standard was also on these slips so that it was presented with each level of the test-indexes. These slips were then attached (stapled) to the 8½" x 11" sheets of paper so that after making a judgment the J could turn the slip and the next level to be judged would be exposed. This procedure was employed in an attempt to ensure that each judgment was independent of the preceding judgment.

The last twelve Js were presented with the following instructions on the last page of the folder: "When the temperature was 101.2° your degree of concern was arbitrarily set at 50. For the diagnostic tests below would you please indicate the value or level that would be necessary for your degree of concern to be also 50."

CHAPTER III

RESULTS

This study was exploratory and descriptive in purpose, and as such it was designed to determine whether or not lawful relationships exist between physicians' judgments pertaining to degree of concern and the results from various numerical medical information indices. That is, were relationships existant, and if so, how could the nature of the relationships best be described. In determining the former there would necessarily have to be a relatively high degree of consensual agreement or reliability between the Js with the various judgmental continua.

For several of the laboratory tests (white blood count, red blood count, specific gravity, pulse) values below and above normal limits were presented to the Js since deviations in either direction are frequently seen in clinical practice. It was decided for each of these tests that all values below normal would best be considered as one set of data, and that values above normal would best be considered as a second set of data. In these instances comparable but separate analyses were conducted with both sets of data. It was believed that all computations should be conducted both with the predetermined standard and without the standard. This was done because, in several instances, it became apparent that the predetermined standard differed quite markedly from a possible intrinsic standard employed by the Js.

Judgmental Reliability

Kendall's coefficient of concordance (*W*), corrected for ties, was utilized in determining the presence and extent of inter-judge reliability. Coefficients were computed for each of the scales, and the significance of each *W* was determined (Siegel, 1956). As can be seen from Table 1, all of these *W* values were significant well beyond the .001 level of significance. This indicates that there was a high degree of consistency or reliability in judgment from one *J* to another. This was true for all tests-indices. Having demonstrated that physicians can reliably judge varying results of clinical laboratory indices, with respect to their concern for an individual's well-being, the formulation of those models which best describe the nature of these relationships was undertaken.

Scale Values

Scale values of concern for each of the nine numerical medical indices were the geometric means of the numbers assigned to the stimuli by the *J*s. When the magnitude scales (degree of concern) were plotted against their respective stimulus metrics (laboratory test results), marked curvilinear relationships were observed with five of the indices. These indices were: white blood count, red blood count, pulse, temperature, and specific gravity. Figures 1a through 5a graphically depict these relationships. When these same values were then graphed with log-log coordinates the curvilinearity was rectified so that straight lines offered reasonably good approximations of the observed trends.

TABLE 1

VALUES FOR COEFFICIENTS OF CONCORDANCE,
S, K, N, AND SIGNIFICANCE LEVELS

Clinical-Laboratory Index	W	s	k,n	P
Specific Gravity*				
Above Normal	.344	4,144.30	26**,5	<.001
Below Normal	.638	2,455.42	26, 4	<.001
Red Blood Count				
Above Normal	.673	3,866.00	26, 4	<.001
Below Normal	.922	3,306.04	27, 4	<.001
White Blood Count				
Above Normal	.958	6,707.34	27, 5	<.001
Below Normal	.952	3,643.54	27, 4	<.001
Pulse				
Above Normal	.903	6,612.08	27, 4	<.001
Below Normal	.818	1,322.43	27, 3	<.001
Diastolic Blood Pressure***	1.000			
Systolic Blood Pressure	.865	24,657.10	27, 8	<.001
Sugar	.922	2,835.00	26, 4	<.001
Protein	.956	3,227.50	27, 4	<.001
Temperature	.908	11,267.83	27, 6	<.001

*For specific gravity values below normal range from 1.000 to 1.016, and values above normal range from 1.016 to 1.040; for RBC the below normal range is from 1.0 to 5.5 million per cu. mm., and above average range was from 5.5 to 13.0 million per cu. mm.; for WBC the below normal range was from 750 to 7,500 per cu. mm., and above average from 7,500 to 200,000 per cu. mm.; for pulse the below normal range went from 40 to 68 per minute, and above average range from 68 to 201 per minute.

**In four instances one J did not make a judgment for one level of the index, therefore, k was equal to 26 rather than 27.

***χ^2test was the appropriate significance test since n = 8 which was larger than tabled values for ($\chi^2 = 216$, df = 8, $p < .001$).

Such rectifications would suggest that power courves might be descriptive of the psychophysical relationships that exist. Power curves were fitted using the least squares methodology.

Product-moment correlations, utilizing logarithmic values, were also computed, both with and without the standard included in the bivi-arate set, between degree of concern scale values and stimulus values to determine the degree of relationship present between these two metric measures. These values, along with the power function exponents for these five scales are found in Table 2.

TABLE 2

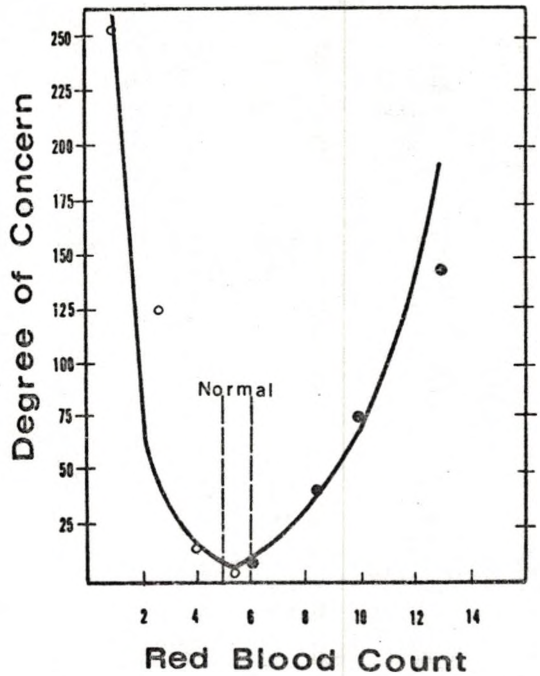
POWER FUNCTION EXPONENTS FOR WBC, RBC, PULSE, TEMPERATURE, SPECIFIC GRAVITY, AND CORRELATIONS OF DEGREE OF CONCERN WITH STIMULUS METRIC

Clinical-Laboratory Index	With Standard			Without Standard		
	Exponent	r	df	Exponent	r	df
Specific Gravity						
Below Normal	-117.72	-.793	3	-154.60	-.983**	2
Above Normal	-------	-----	-	71.07	.977†	3
Red Blood Count						
Below Normal	-------	-----	-	-2.34	.929*	2
Above Normal	3.82	.921†	4	4.12	.978†	3
White Blood Count						
Below Normal	-------	-----	-	-2.21	-.950**	3
Above Normal	1.30	.976‡	4	1.46	.998‡	4
Pulse						
Below Normal	-------	-----	-	-5.25	-.978	1
Above Normal	3.65	.906†	5	4.86	.976†	4
Temperature	37.66	.976‡	5	42.17	.998‡	4

*p. < .10 †p. < .01
**p. < .02 ‡p. < .001

Figure 1a. Relationship between magnitude estimation of degree of concern and red blood count.

Figure 1b. Relationship between log degree of concern and log red blood count. (Data are from Figure 1a.)


Degree of Concern
250
225
200
175
150
125
100
75
50
25
Normal
2
4
6
8
10
12
14
Red Blood Count

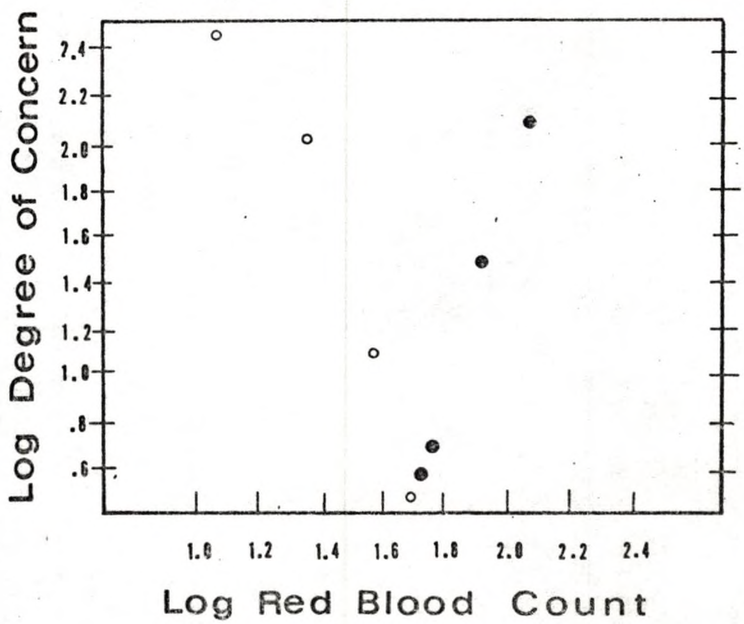
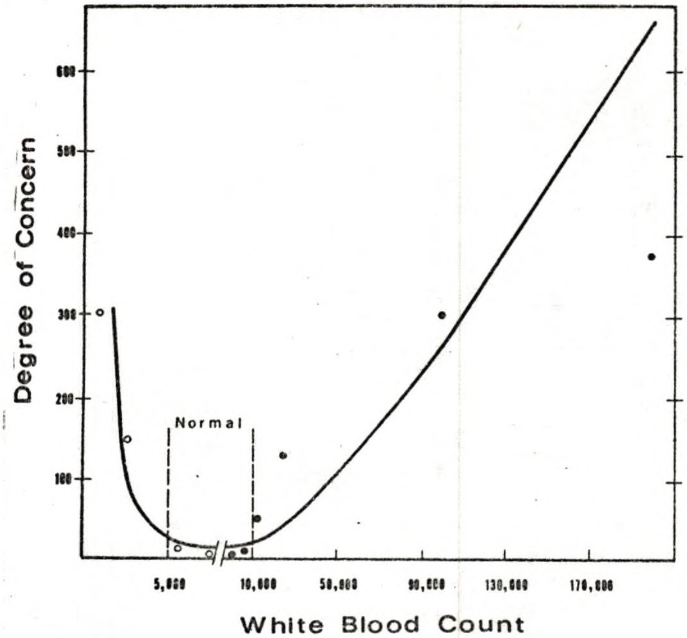

Log Degree of Concern
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Log Red Blood Count

Figure 2a. Relation between magnitude estimation of degree of concern and white blood count.

Figure 2b. Relation between log degree of concern and log white blood count. (Data are from Figure 2a.)


Degree of Concern
600
500
400
300
200
100
Normal
5,000
10,000
50,000
90,000
130,000
170,000
White Blood Count

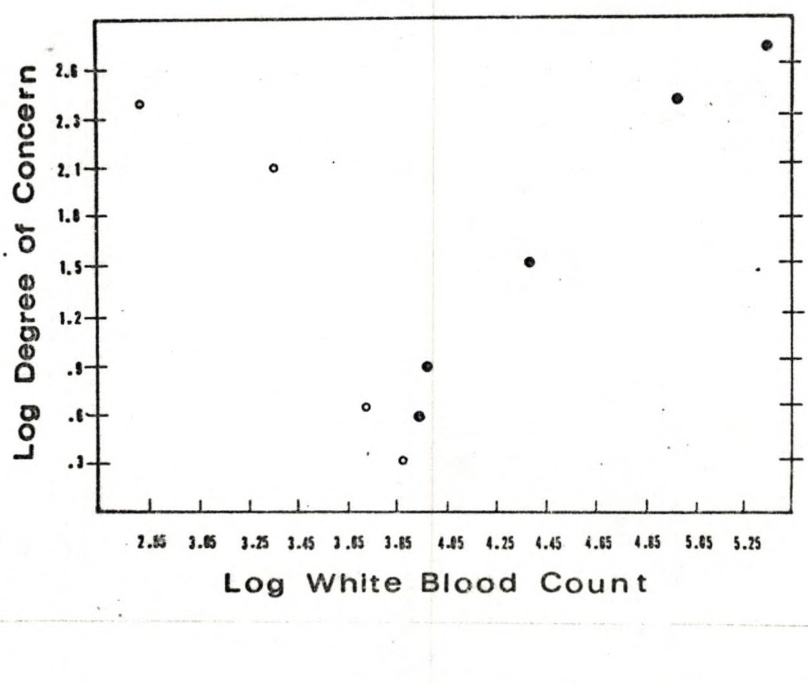
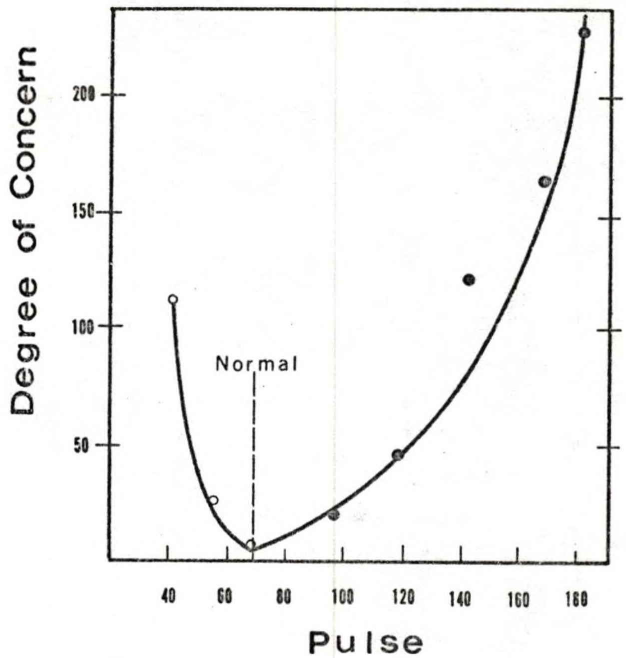

Log Degree of Concern
2.6
2.3
2.1
1.8
1.5
1.2
.9
.6
.3
2.85
3.05
3.25
3.45
3.65
3.85
4.05
4.25
4.45
4.65
4.85
5.05
5.25
Log White Blood Count

Figure 3a. Relation between magnitude estimation of degree of concern and pulse.

Figure 3b. Relation between log degree of concern and log pulse. (Data are from Figure 3a.)


Degree of Concern
200
150
100
50
Normal
40
60
80
100
120
140
160
180
Pulse

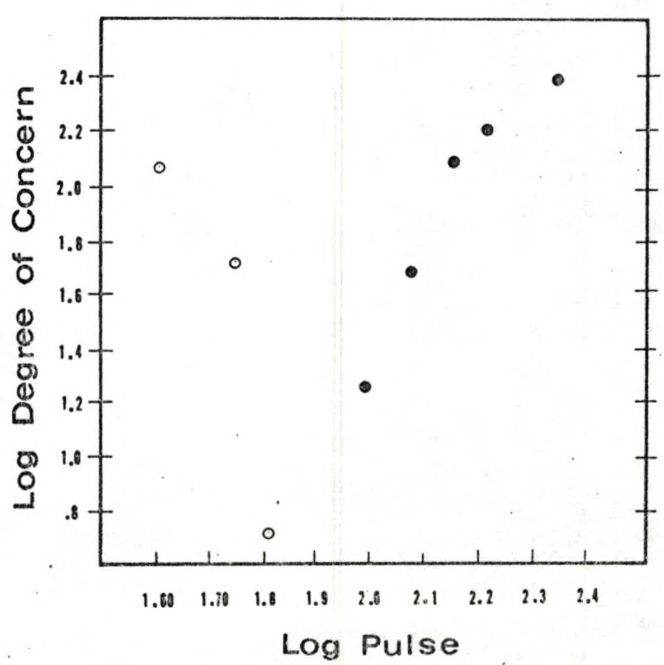
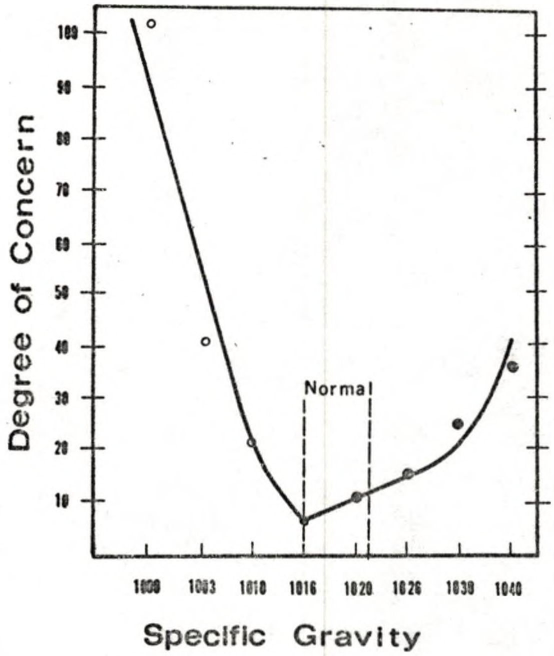

Log Degree of Concern
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
.8
1.60
1.70
1.8
1.9
2.0
2.1
2.2
2.3
2.4
Log Pulse

Figure 4a. Relation between magnitude estimation of degree of concern and specific gravity.

Figure 4b. Relation between log degree of concern and log specific gravity. (Data are from Figure 4a.)


Degree of Concern
100
90
80
70
60
50
40
30
20
10
Normal
1000
1003
1010
1016
1020
1026
1030
1040
Specific Gravity

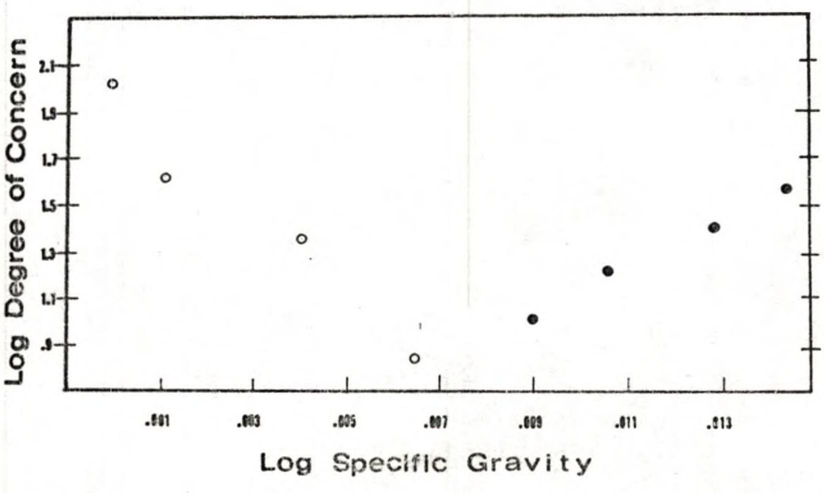
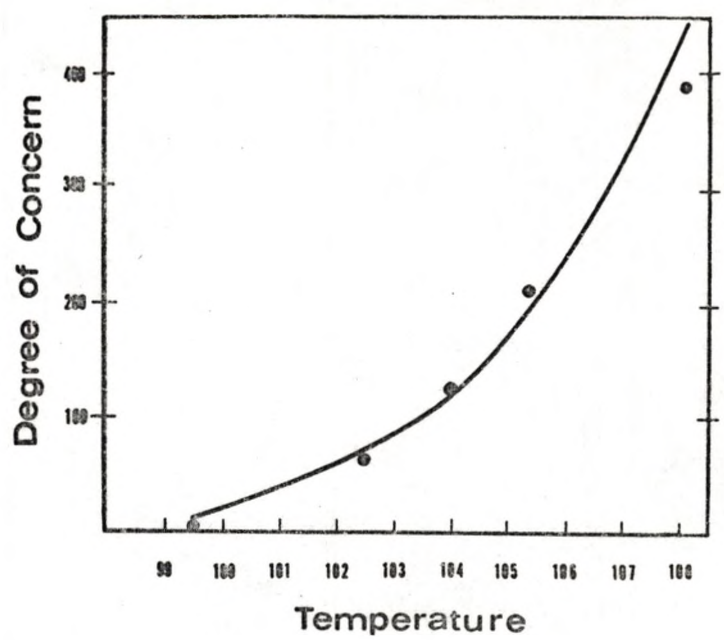

Log Degree of Concern
2.1
1.9
1.7
1.5
1.3
1.1
.9
.001
.003
.005
.007
.009
.011
.013
Log Specific Gravity

Figure 5a. Relation between magnitude estimation of degree of concern and temperature.

Figure 5b. Relation between log degree of concern and log temperature. (Data are from Figure 5a).


Degree of Concern
400
300
200
100
99
100
101
102
103
104
105
106
107
108
Temperature

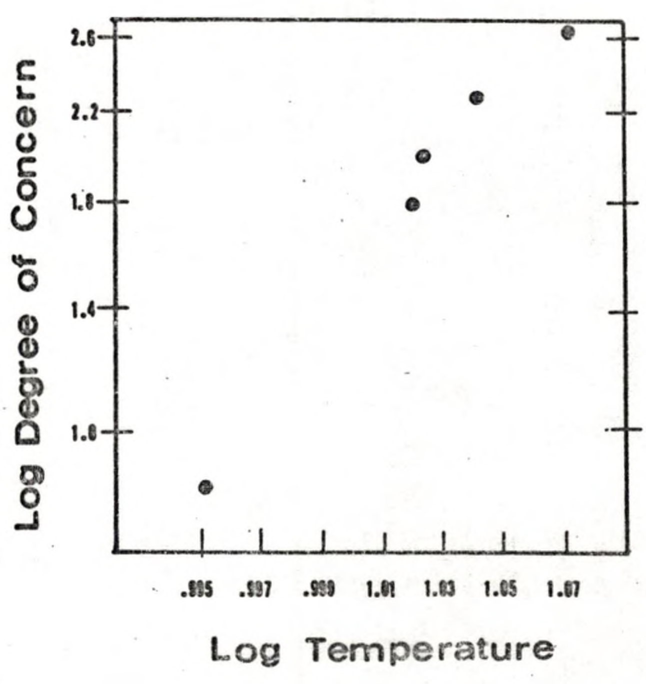

Log Degree of Concern
2.6
2.2
1.8
1.4
1.0
.995
.997
.999
1.01
1.03
1.05
1.07
Log Temperature

TABLE 3

SLOPES OF THE LINES FOR SYSTOLIC AND DIASTOLIC BLOOD PRESSURE, SUGAR, AND PROTEIN, AND CORRELATIONS OF DEGREE OF CONCERN WITH STIMULUS METRIC

Clinical-Laboratory Index	With Standard			Without Standard		
	Slope	r	df	Slope	r	df
Systolic Blood Pressure	1.45	.963*	8	1.45	.957*	7
Diastolic Blood Pressure	1.94	.993*	6	2.12	.994*	5
Sugar	42.51	.996*	3	----	-----	-
Protein	47.98	.998*	3	----	-----	-

*$P. < .001$

relationship are not applicable with the present data for several reasons. First, the degree of concern scale values were calculated as geometric means rather than as arithmetic means. Secondly, there was some heterogeneity of variances (0.166 - 0.939 in logs) which would violate the homoscedasticity assumption necessary for these traditional goodness of fit tests (Lewis, 1960). Thirdly, these two statistical analyses require that observations be independent of each other. The observations in this investigation were not independent since each J made judgments with respect to all levels on the information indices.

In lieu then, of a fully appropriate statistical technique for testing goodness of fit, product-moment correlations were computed between magnitude estimation scale values and predicted magnitude scale values derived from the straight line functions and power functions.

Figure 6. Relationship between magnitude estimation of degree of concern and systolic blood pressure.

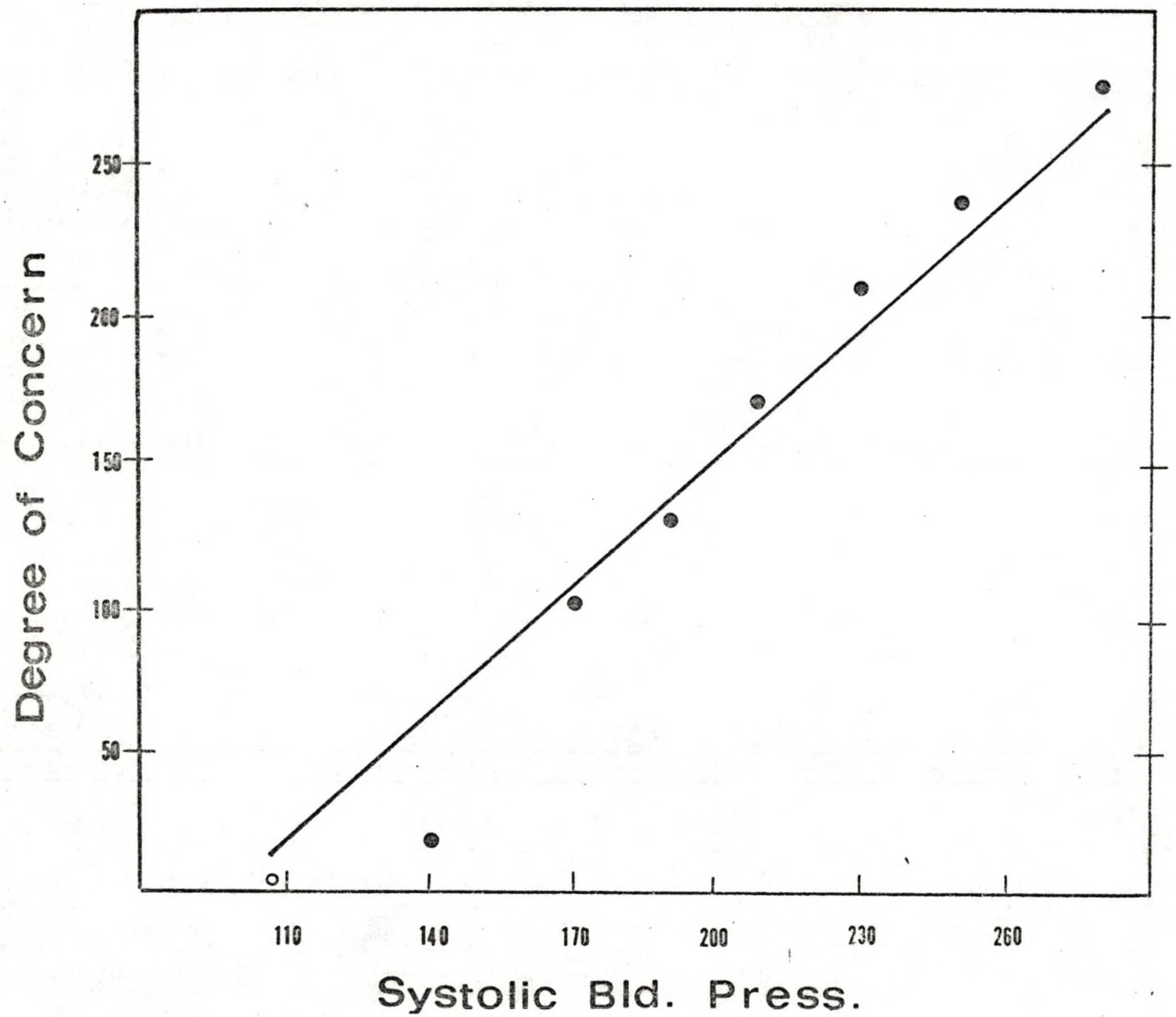

Degree of Concern
250
200
150
100
50
110
140
170
200
230
260
Systolic Bld. Press.

Figure 7. Relationship between magnitude estimation of degree of concern and diastolic blood pressure.

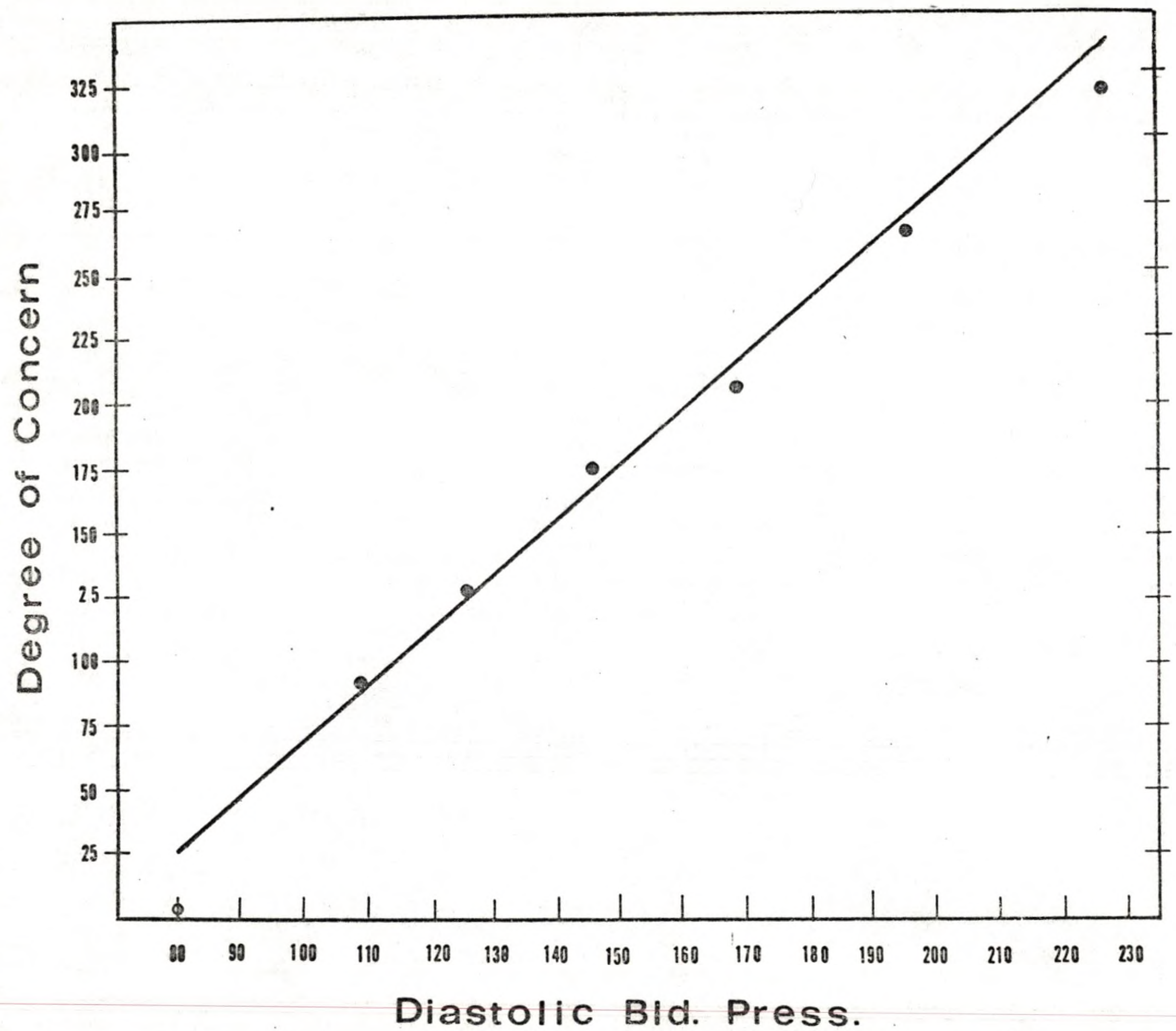
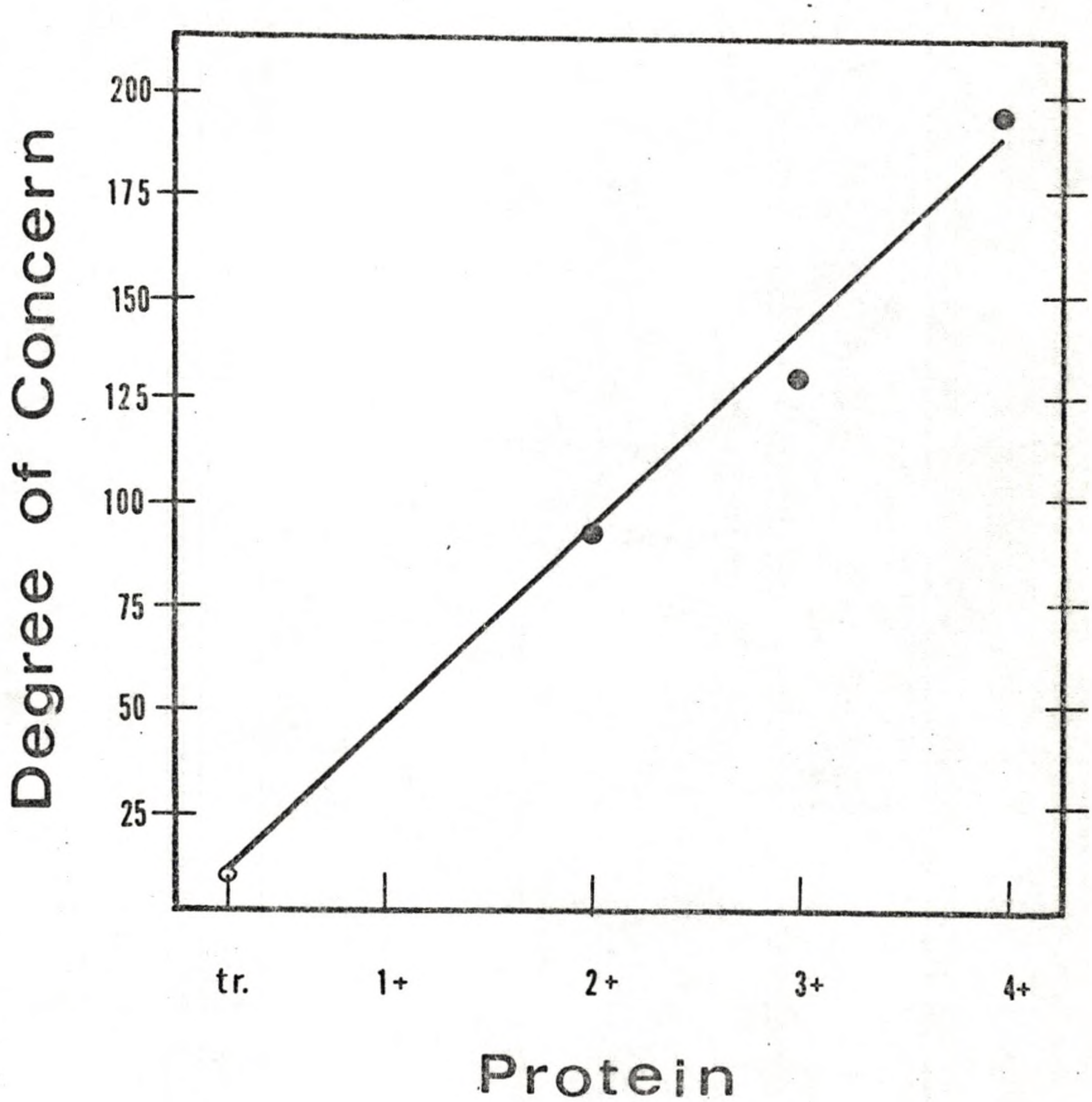

Degree of Concern
325
300
275
250
225
200
175
150
2 5
100
75
50
25
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
Diastolic Bld. Press.

Figure 8. Relationship between magnitude estimation of degree of concern and protein.



Degree of Concern
200
175
150
125
100
75
50
25
tr.
1+
2+
3+
4+
Protein

Figure 9. Relationship between magnitude estimation of degree of concern and sugar.

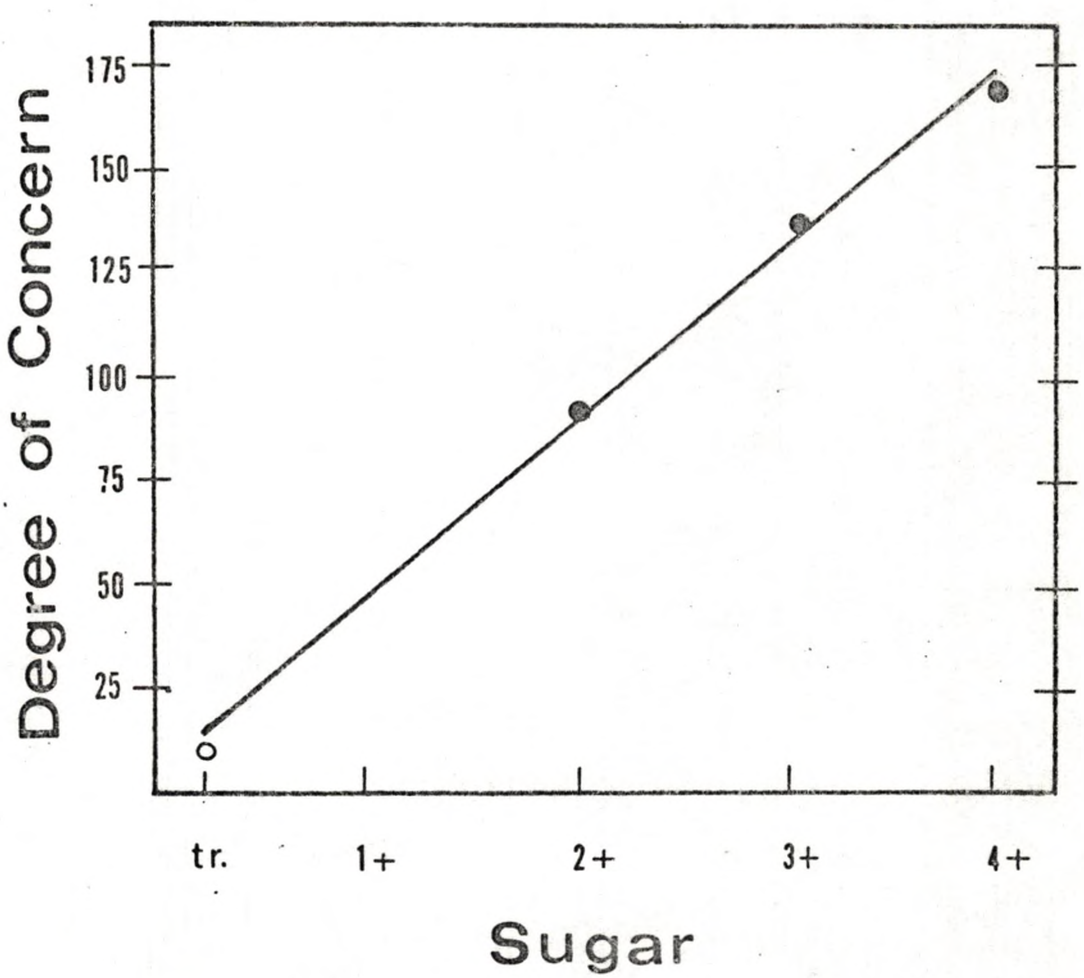
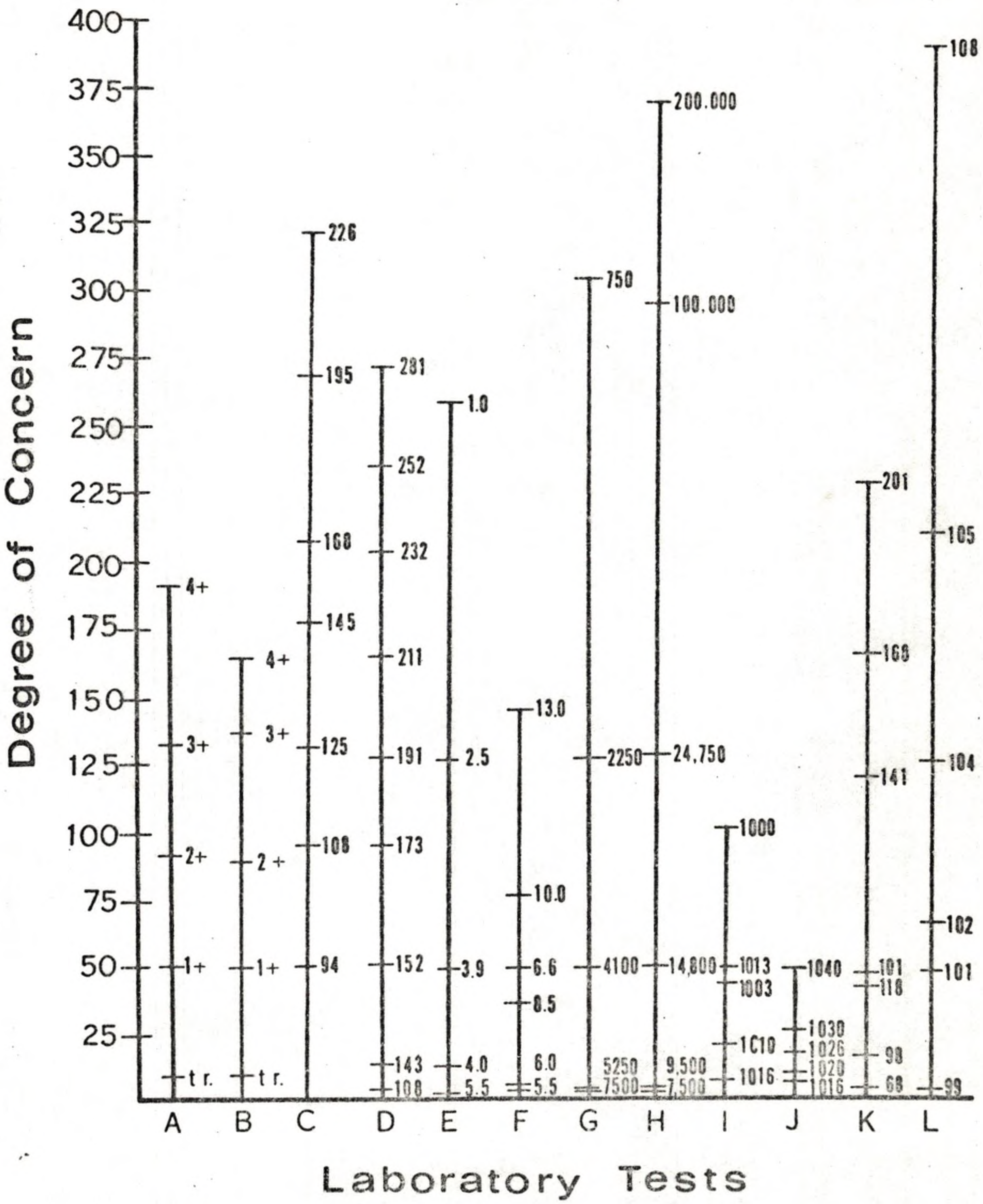

Degree of Concern
175
150
125
100
75
50
25
tr.
1+
2+
3+
4+
Sugar

Figure 10. Relationship between laboratory tests when equated for a subjective concern value of 50. Codings for the laboratory tests are: (A) sugar, (B) protein, (C) diastolic blood pressure, (D) systolic blood pressure, (E) red blood count (below normal), (F) red blood count (above normal), (G) white blood count (below normal), (H) white blood count (above normal), (I) specific gravity (below normal), (J) specific gravity (above normal), (K) pulse, (L) temperature.


Degree of Concern
400
375
350
325
300
275
250
225
200
175
150
125
100
75
50
25
Laboratory Tests
A
B
C
D
E
F
G
H
I
J
K
L
4+
3+
2+
1+
t r.
4+
3+
2 +
1+
t r.
226
195
168
145
125
108
94
281
252
232
211
191
173
152
143
1.0
2.5
3.9
4.0
5.5
13.0
10.0
6.6
8.5
6.0
5.5
750
2250
4100
5250
7500
200.000
100.000
24,750
14,800
9,500
7,500
1000
1013
1003
1016
1040
1030
1026
1020
1016
201
168
141
101
118
98
68
108
105
104
102
101
99

These values are presented in Table 4. The correlation coefficients may be used to show the quality of fit illustrated by the straight line functions and power functions graphically depicted in Figures 1a through 9. Since all correlations, except two, are significant beyond the .05 level they suggest that the depicted functions may provide appropriate descriptions. The exceptions were pulse (without the logarithmically selected standard), and specific gravity (with the logarithmically selected standard). These exceptions are the same found previously when the correlations between stimulus magnitude values (cf. Table 2) were discussed. Again, the one _df_ allowed for the pulse scale, and perhaps, the inappropriateness of the predetermined standard utilized for specific gravity is believed to account for the nonsignificance.

For each clinical-laboratory index, standard deviations for magnitude estimations (without the logarithmically determined standards) were computed for each stimulus value. Product-moment correlations between the logarithmically transformed scale values and their standard deviations were computed to test the proposition, as stated in Ekman's law (cf. Stevens, 1966), that judgmental variability, with respect to magnitude estimations, increases with subjective magnitude on prothetic continua. These values can be found in Table 5.

TABLE 4

CORRELATIONS BETWEEN OBSERVED AND PREDICTED MAGNITUDE ESTIMATION SCALE VALUES

Clinical-Laboratory Index	With Standard r	With Standard df	Without Standard r	Without Standard df
Pulse				
Below Normal	------	-	.981	1
Above Normal	.867**	5	.967†	4
Specific Gravity				
Below Normal	.789	3	.986**	2
Above Normal	------	-	.983†	3
Red Blood Count				
Below Normal	------	-	.932*	2
Above Normal	.915**	4	.993*	3
White Blood Count				
Below Normal	------	-	.950**	3
Above Normal	.882**	4	.891**	4
Temperature	.991†	5	.983**	4
Systolic Blood Pressure	.899‡	8	.961‡	7
Diastolic Blood Pressure	.991‡	6	.961‡	5
Sugar	.995‡	3	------	-
Protein	.996‡	3	------	-

*p. < .10   †p. < .01

**p. < .02   ‡p. < .001

TABLE 5

CORRELATIONS OF LOGARITHMICALLY TRANSFORMED ESTIMATION SCALE VALUES WITH STANDARD DEVIATIONS OF THE SCALE VALUES

Clinical-Laboratory Index	r	df	Clinical-Laboratory Index	r	df
White Blood Count			Specific Gravity		
Below Normal	-.877	2	Below Normal	-.938*	2
Above Normal	-.866**	4	Above Normal	-.923**	3
Red Blood Count			Temperature	-.183	4
Below Normal	-.950**	2			
Above Normal	-.991‡	3	Pulse		
			Below Normal	-.932	1
Systolic Blood Pressure	-.709**	6	Above Normal	-.939***	3
			Sugar	-.978***	2
Diastolic Blood Pressure	-.872***	5	Protein	-.945*	2

*p. < .10 ***p. < .02
**p. < .05 ‡p. < .001

It is readily observed that the relationships between concern scale values and variability (standard deviations) were directionally consistent. All correlational values were negative. Only three correlations failed to reach at least the .10 level of significance. Those three correlations were: WBC (for scale values below normal), pulse (only for the values below normal), and temperature. For WBC it was found that the logarithmically selected standard was considered to be within normal limits by the Js. This resulted in an increase in judgmental variability about those stimulus values within the accepted normal limits (see Appendix). Correlations computed for pulse values below normal were again found to be nonsignificant, as they were throughout this investigation. Since all correlations with this scale were computed on just three points, only one df was allowed in significance tests. This restrictive factor was believed to be the major reason for the nonsignificant findings. Should more points in this pulse range be added it would be expected that statistically significant findings would be obtained. In the case of temperature, the nonsignificant correlation appears to be a function of judgmental confusion. For example, the stimulus which had the largest standard deviation on this scale was 99^{o}, only .4 of a degree above normal (98.6^{o}). It would seem that perhaps the Js found it difficult or confusing to determine subjective concern for a stimulus so close to the value generally considered as normal.

Ekman's law holds that judgmental variability, in subjective unit measurement, grows as a linear function of subjective magnitude, i.e., as subjective magnitude increases judgmental variability also increases. The present investigation does not permit a direct test of this law. For a direct test of this law it would be necessary to ascertain subjective magnitude for the information indices. The present study scaled degree of concern rather than subjective magnitude. However, the correlations found in Table 5 indicate that the subjective concern scale variabilities were linearly related to the objective magnitudes of the stimuli. That is, as the stimuli became "more pathologic", judgmental variability decreased. It would not seem unwarranted to assume that there is an intermediate scale between subjective concern and the stimulus metric. Guilford

(1954) has argued for an intervening judgmental continuum. In the context of the present investigation, an intermediate scale could be one of subjective magnitude.

On the assumption of such an intermediate scale, perhaps an indirect test of Ekman's law can be offered. In the present inquiry, judgmental variability was less for the more pathologic stimuli, and greater for those stimuli close to and within accepted normal limits. It is suggested that subjective magnitude (in terms of a scale producing response) and subjective concerned scaled in this investigation (as a precursor to a response) are related. Following this reasonable suggestion, it is not inconceivable that with "pathologic" stimuli this created less judgmental variability since these stimuli may have served as a mandate for imperative diagnostic and subsequent treatment response. The less pathologic stimuli perhaps do not subjectively carry such a mandatory treatment or action message. For example, a temperature of 108^{o} would seem to demand and compel immediate action from all physicians. Whereas a temperature of 101^{o} would not seem to require such an immediate treatment response. The physician, therefore may wish to consider a variety of possible etiological factors and treatment regimens. His subjective concern then, as a precursor to a response, may reflect more variability (of possible treatment action responses) at the less pathologic levels.

For all laboratory test indices, except temperature, the last twelve <u>J</u>s to participate in the study were asked to indicate the value or level that a particular test-index would have to be in order to produce a degree of concern having the value 50. This inquiry was accomplished for three purposes. First, to provide a rough "check" on the appropriateness of the standard stimuli utilized in the present investigation. Secondly, the results can provide some basis for future research in terms of providing appropriate "equal-concern" standards. Thirdly, the results can provide some tentative information about degree of concern across scales, and to indicate those scales which are the most or least potent in terms of eliciting concern. In several instances (e.g., specific gravity, pulse, RBC) the logarithmically determined standard was found to be quite inappropriate or disparate from the values offered by the Js themselves as being necessary for them to feel a degree of concern equivalent to 50. The effect of this disparateness was seemingly to add more "noise" to the judgmental process through confusion. The result of such confusion was believed to be expressed in increased judgmental variability in the judgment making process.

The Js were asked to supply a level for each test-index that would produce a subjective concern of 50. The means of the responses to this request are given in Figure 10 at the concern value of 50. The remaining values are those obtained from the Js after the standard had been established, i.e., they are the obtained magnitude estimation scale values. Only three "reversals" are seen, i.e., where higher or more "pathologic" stimuli (index levels) had scale values smaller than the stimuli the Js believed to be associated with a subjective concern of 50. These inconsistencies are found with pulse, specific gravity, and RBC. It is seen that these "reversals" or inconsistencies have concern values only slightly less than 50, and the "reversals" involve only small differences. It is of interest to note however, that these occur on the scales where the standard was felt (as reported by several _J_s) to be somewhat inappropriate. Such reversals may well reflect the addition of judgmental "noise" into the measurement system.

CHAPTER IV

DISCUSSION

The present investigation was designed to describe the relationship between physicians' subjective degree of concern for a specified patient's well-being when the results from nine frequently utilized numerical test-indices were varied systematically and independently. In general, the results illustrate that with five of these medical indices the relationships were curvilinear when degree of concern was plotted against the appropriate stimulus metric. It was determined that a power function model offered an appropriate description of these bivariate data. For the remaining four medical indices the relationships, when degree of concern was plotted against the stimulus continua, were markedly linear in nature.

Judgmental Reliability

As indicated in Table 1, there was a pronounced degree of concordance or agreement (inter-judge reliability) between Js for each of the nine clinical indices. This might be expected since the medical information measures utilized in this investigation were selected because they are frequently considered an integral part of the diagnostic process. They are therefore, utilized frequently by physicians. The nine measures may then be considered as routine in most medical examinations. The ordinal or interval numerical nature of the stimuli also may have facilitated the high judgmental concordance. For example, a temperature of 105.4^{o} is clearly seen to be higher than one of 103.5^{o}. Another factor which also may have promoted the observed high degree of consensus was the seniority of the Js. They had considerable clinical experience, an average of twenty-one years in practice. It would be expected that their judgments were seasoned and would reflect a high degree of consensus.

However, the high degree of judgmental reliability obtained is notable when two other aspects are considered. The Js represented a diverse group of medical specialities, i.e., physicians in six different specialities interpreted the numerical indices. Also, there appears to be a paucity of medical literature which deals with relationships between numerical information indices and concepts such as concern, suspicion, or prompt action on the part of the physician. The literature reviewed in the present investigation, for the most part, used such general terms as "an elevation of systolic pressure," or a "pronounced raise...." Perhaps, the high judgmental consensus manifested was the result of the physicians' common experiental programs such as internships, residencies, etc., as well as the factors discussed in the previous paragraph. Several Js, when questioned on this matter, could not recall any formal didactic training experiences in which specific levels of medical indices were linked to concepts such as concern, suspicion, or vigilance.

Judgmental reliability is a necessary prerequisite to any psychological scaling efforts, and it is the foundation of all acceptable scaling methodologies. As Underwood (1957, p.22) notes, "If our response measure is not reliable, no further investigative procedures should be undertaken. Science attempts to discover and understand reproducible phenomena; lack of reliability in our attempts at measurement precludes this reproducibility." The present inquiry observed high judgmental concordance. Therefore further analyses appeared justifiable. These procedures seemed to explicate the quantitative nature of the observed relationships between degree of concern and medical information indices.

Linear and Curvilinear Functions

The relationships between subjective concern and four of the information indices were regarded to be linear functions. Of these four, two (protein and sugar) are typically reported in terms of a six-point category scale (e.g., 0, trace, 1+, 2+, 3+, 4+). It may be that the underlying subjective concern continua for these two indices, as well as those of systolic and diastolic blood pressure, are in fact metathetic. For the other five information indices, curvilinear relationships between degree of concern and test-indices were observed. Perhaps the underlying subjective continua for these five may be prothetic in character. However, a determination as to whether the continua underlying protein, sugar, and both of the blood pressure measures are metathetic or prothetic is difficult since this study was not specifically designed to test this distinction. For a validation of this hypothesis further investigations in which category ratings are also obtained on these four indices would have to be conducted. The resultant category scales could then be plotted against magnitude estimation scales such as those developed in this investigation. If relationships between the two were found to be concave downward, one could more reasonably advance the hypothesis that the judgmental continua underlying these five indices are prothetic in character. If the interscale relationships were linear, it would be suggestive of a metathetic continua.

It is interesting to note that the linear relationships between subjective concern and information indices on four of the scales were quite high (cf. Table 3). Inspection of the graphic depiction for these four scales (Figures 6-9) shows, in general, that departures from the lines of best fit are rather minor. In fact, the departures appear to be less than the ones observed for some of the five scales characterized by curvilinear functions (cf. Figures 1a-5b). In addition, the predetermined standard appeared appropriate for the four linear scales since the correlations computed with and without the standard are very similar. This was not true for those five scales characterized by curvilinear functions. In every case, the correlations between subjective concern and the stimulus metric were higher when computed without the standard. This perhaps suggests that the standards employed for these five scales were "misfits."

For the five clinical-laboratory measures which could be regarded as related to subjective concern by power functions, the exponents (of the power functions) were all greater than unity. This may be interpreted to mean that doubling of the stimulus metric (numerical index) results in more than a doubling in the related psychological intensity scale (degree of concern). This is also the case for the four linear relationships observed. For example, a RBC level of six million created a subjective concern of about 14, and a RBC level of twelve million resulted in an almost tenfold increase in degree of concern. Physicians have certainly acquired or received considerable common "input" information regarding the significance of the different indices as indicators of health status for patients. Such information apparently is not derived from formal didactic experiences but seems to be accumulated through clinical experiences.

Standard Selection

Stevens (1957, p.167) notes that "...when people try to describe a sensation in quantitative terms they face a difficult task, and factors that affect the outcome are numerous and subtle." Two factors in the present investigation which may have had an affect upon the judgment process should be considered. These factors are the appropriateness of the utilized standard, and the numerical value of the assigned modulus.

For specific gravity, pulse, and RBC, it was rather apparent that the logarithmically selected standard was disparate from that used intrinsically by the physician-judges. The net effect of this disparatness was reflected in the standard deviations associated with the subjective scale values of those stimuli closest to these standards. Judgmental variability was quite marked for these three indices. This notable variability was reflected in the large standard deviations associated with the three information measures (see Appendix). This increased variability illustrates the additional "noise" added to the measurement process when inappropriate standards are utilized. For scaling methodology in general, these results would suggest that careful attention must be given to standard selection. Inappropriate standards seem to be a source of extraneous contamination in terms of creating additional judgmental variability. Standards which closely resemble the intrinsic or "natural" standards of judges would be expected to reduce judgmental variability and increase measurement precision.

With regard to the standards selected for specific gravity and pulse, several physicians volunteered that they considered the standards to be within normal limits. It was therefore quite difficult for them to make proportional judgments of concern relative to a normal standard. For specific gravity, three Js offered that this laboratory test becomes meaningful for them only when the test is repeated several times under conditions which allow them to regulate food and fluid intake. Additionally, a further methodological note concerns the manner in which the specific gravity values were presented to the Js. One physician noted that results are usually reported in four digits whereas the present investigation employed five digits. This may have caused indeterminable confusion in the judgments. All three of these factors were believed to be associated with increased judgmental variability with respect to the specific gravity scale.

A similar effect was observed on the RBC scale, as the judgmental variability was pronounced. Once again the predetermined standard (7.0) proved somewhat disparate from that suggested by the _J_s (6.5). In contrast with the evaluation pertaining to specific gravity and pulse discussed in the preceding paragraph, where the predetermined standard was felt to be within normal limits, the RBC standard was felt by some to be "too pathologic". Several _J_s stated that it was too far above normal limits to warrant a subjective concern value of only 50. The judgmental effect of this disparateness, however, was believed to be similar to that observed for specific gravity and pulse, i.e., it increased judgmental variability.

The second factor which may have had an effect upon the judgmental process was the numerical modulus value for the standard. Poulton (1968) presents convincing evidence that the numerical value selected as an assigned modulus may affect the slopes of psychophysical power function exponents. Specifically, a modulus near the extreme ends of a possible range of numbers appears to create less steep slopes for power functions.

In the present inquiry, 50 was used as the numerical value for the assigned modulus. It was believed that the judges would have considerable or sufficient freedom to select numbers on either side of this modulus to reflect their degree of concern. However, one judge noted that he would not use a number (in making magnitude estimations) larger than 200. He utilized this self-imposed upper limit despite encouragement and assurances from the investigator that he could assign any number he wished to express his degree of concern. The effect of this self-imposed restriction was to reduce the discrimination of his judgments, especially at the more pathologic levels of the information indices.

This example clearly illustrates one of the difficulties encountered in ratio estimation methods, that is, idiosyncratic number usage by judges. For example, the number 50 may have appeared (subjectively) to represent a high degree of concern to one judge and a lesser degree of concern to another judge. This subjective difference could be attributable to either personal perceptions of the number or cultural factors which could also affect individual number perception. Ekman and Sjoberg (1965) note that the theory inherent in the use of these methods implicitly assumes that the judge utilizes numbers in the same way as the psychologist or mathematician. This would be difficult to prove. Nevertheless, with only the one exception noted, no direct evidence was found to indicate that idiosyncratic number usage had a marked effect upon the observed bivariate relationships. Future investigations similar to the present one may wish to consider this factor since it may effect judgmental variability and psychophysical power function slopes.

Inter-Scale Comparisons

Little is known about inter-scale comparability. That is, there is little evidence concerning how pathologic levels on one index compare to pathologic levels on another information index with regard to creating subjective concern. For example, does a RBC level of 10.0 million per cu. mm. arouse more or less subjective concern than a systolic blood pressure of 180? The physician-judges were asked to equate test-index levels for a subjective concern value of 50. It was believed that these comparisons would possibly provide some suggestions upon which to base inter-scale comparisons.

Values for subjective degree of concern on each medical index, when the level of the tests were equated for a subjective concern value of 50, are seen in Figure 10. Interpretation of the figure reveals that, with reference to perceived pathology, the highest level of the temperature (108°) measure used elicited the highest degree of scaled concern when compared to the other clinical indices. It is also seen that RBC and both of the blood pressure indices, when pathologic levels were reported, were subjectively more potent (in creating concern) when compared to the other information indices. The upper limites of these indices seemingly suggest greater pathology as compared to the other information indices. With only one exception, levels creating a degree of subjective concern equal to 250 or greater were associated with the temperature, WBC, and both of the blood pressure indices. This one exception was a RBC stimulus level of 1.0 million per cu. mm. This level seemed to be associated with more than twice the subjective concern manifested for the next closest stimulus level, 2.5 million per cu. mm. In general, these results indicate that the extreme pathologic values on temperature, WBC, and both blood pressure indices elicited the greatest subjective concern from the physicians sampled. Certainly, these four indices appear to possess the greatest potential for eliciting medical attention (concern) when pathologic levels are noted. Perhaps upper and lower level results from these four tests are given the most considered medical attention. This is, of course, relative to the 35 year old patient specified in the instructions. It would be most interesting to determine if these indices levels are so perceived by other physicians, and whether or not scaled concern would remain similar with other types of patients.

Specific gravity, levels above and below normal, failed to elicit the extremes of subjective concern. As seen in Figure 10, only one such value elicited a concern greater than 50, and that level was 1.000. Such a specific gravity is not biologically possible as it is the specific gravity of distilled water. Perhaps the reason it elicited the scaled degree of concern that it did was due to confusion caused by bewilderment of how to respond to this highly unlikely stimulus. The notable low capacity to elicit subjective concern on this particular laboratory index may reflect that the physician-judges considered it to be of limited informational value, comparatively speaking, when considering pathological processes. However, several other factors should be reiterated about this scale. The reader will recall that the standard employed was found to be inappropriate. Also, the test results were presented in five digits rather than the customary four. Furthermore, several Js noted they felt results from this index to be meaningful only under highly controlled conditions, i.e., repeated measures, knowledge of fluid intake, etc. Therefore, whether the low subjective concern potency is attributable to these methodological considerations or to a genuine lack of significance (relative to the other indices) is indeterminable at this time.

The other six numerical information indices appeared to be of approximately equal puissance in terms of their potential to create subjective medical concern.

Future research may well wish to consider the data shown in Figure 10 for yet another reason. The values associated with a subjective concern value of 50 would seem to provide appropriate standards for future investigations employing direct estimation methodologies for these kinds of stimuli. These levels of equal concern were suggested by the physician-judges with very little variability in their suggestions. Use of a predetermined standard which closely approximates the intrinsic or "natural" standard utilized by judges should eliminate one possible source of extraneous judgmental variability. In the present investigation for example, those indices which had the least judgmental variability associated with them were also those indices for which the predetermined standard closely approximated the Js' "natural" standard. Greater judgmental variability or "noise" was associated with those indices where the predetermined and intrinsic standards were quite disparate. A reduction in judgmental variability is desirable as it increases reliability and therefore increases measurement precision.

Implications and Limitations

This study provided some implications for the direct estimation literature as it represents one of the earliest extensions of these techniques into the area of clinical judgment. Only one published series of studies (Stone, 1966, 1968a, 1968b, 1969; Stone and Skurdal, 1968) has made such a promising and suggestive application of the direct estimation methods with clinical judgmental material. In those investigations, and in the present inquiry, the direct estimation methodologies proved to be valuable tools for elucidating the heretofore rather nebulous and murky area of clinical judgment. It is felt that many previous investigations pertaining to clinical judgment and decision making have been deficient in that the methodologies employed have not been sufficiently sensitive to the judgmental process so that more definitive conclusions could be drawn. This would perhaps explain the many apparently conflicting results. On the basis of this inquiry and on those of Stone, it might be suggested that direct estimation procedures do indeed provide sufficient sensitivity to assist in the clarification of the many enigmatic ambiguities now existent in the clinical judgment literature.

Significant implications can be drawn from this investigation for clinical-medical education. As noted in the Results chapter, several Js could not recall having been exposed to didactic experiences in which "degree of concern" for specified levels of medical indices were expressed or made explicit. The investigator also was unable to find any such reference or even one closely similar to it based on a rather extensive review of medical literature. The high degree of inter-judge agreement indicates that perhaps subjective impressions similar to those scaled in the present investigation are developed intrinsically by physicians. These impressions would appear to be utilized in clinical practice with a high degree of consensus. Based on the results of this inquiry, it is believed that similar subjective scales of concern could be developed so as to provide clinical-medical education with succinct graphic depiction of some of the parameters of seasoned clinical judgment. In the development of these scales attention should be given to several of the methodological suggestions considered in this inquiry. Future investigations might seemingly produce valuable communication vehicles so that the exigencies of the office could be brought into the classroom.

There is nothing short of a myriad of possible extensions of the present inquiry into the general area of clinical-medical judgment. For example, various medical specialities might develop a list of their respective commonly encountered "warning signals" or diagnostic tests. These then could be subjected to direct estimation evaluations. It would also be interesting and perhaps profitable to compare and contrast different medical specialists' interpretations of test results. Other studies could consider interpretations of clinical medical indices by physicians who have had different amounts of clinical experience.

A comment that could be relevant for future investigations, and which points to limitations in the present study should be noted. Future investigations may wish to manipulate simultaneously several information measures which are similar to those in this study. This was not accomplished in the present investigation since it was believed necessary to provide a situation which afforded good experimental controls. For the physician, the diagnostic-judgment process is a dynamic, fluid situation that "changes continuously, and in which the doctor and patient interaction works many ways." (Rimoldi, 1964, p. 328) Because the judgmental situation is dynamic it becomes necessary to attempt to isolate some of its interdependent facets. This is experimentally realized at the expense of sacrificing the "real world" situation. Exploratory studies, to be productive and meaningful, frequently impose laboratory conditions upon the topic under investigation. An attempt to reduce a very complex phenomenon to something more simple is many times the sequence often followed in scientific inquiry. Especially in exploratory inquiries it seems necessary and efficacious to first describe and understand the basic or primary processes. With such understandings, subsequent investigations can better explore the interactions and subtle complexities of the "real world".

Being cognizant of the complex nature of the "real world", it was believed efficacious to impose some experimental controls. In actual clinical practice physicians purport to consider all test results in an interrelated manner. The present investigation's results can offer to future investigators, a sounder knowledge of relevant variables, standards, modulus values for standards, etc.

In the usual clinical situation the physician seldom, if ever, is asked to base his diagnostic-treatment judgment solely on the results of just one clinical-medical index. A number of physician-judges expressed this concern when first asked to participate. One physician summed up this point well with the statement, "This is not the way the real world operates." A study considering such a comment might examine the same or similar variables as employed in this investigation in terms of their possible interactions. If successful, such a study would allow one to determine what degree of subjective concern would be associated with various combinations of medical test-indices. For example, if a temperature of 101° equaled 50 units of concern and a WBC level of 18,000 equaled 75 units of concern, it would then be possible to state that, when these two findings existed simultaneously, the scaled degree of concern would be some different value. Obviously, the additive nature of the subjective concern values could be considered. If it were possible to understand medical indices in this fashion, the implications for clinical-medical education are many.

CHAPTER V

SUMMARY AND CONCLUSIONS

The nature of the present investigation was essentially exploratory and descriptive. Utilizing the direct scaling method of magnitude estimation (with assigned modulus) the investigation attempted to determine if lawful relationships existed between the clinical judgmental variable of "concern" (based on judgments of licensed physicians) and the results from nine numerical medical information indices. A quantitative description of such relationships was also planned.

It was believed that direct estimation methodologies have been shown, through many empirical studies, to be superior to psychophysical models founded on the ideas of Fechner and Thurstone and on the rating scale methods. Previous work has also illustrated that the power law model suggested by S. S. Stevens has provided a meaningful methodology in studying the topic of clinical judgment. Controversial and equivocal suggestions have resulted from prior research. Nevertheless, clinical judgment was seen as bein amenable to scientific scrutiny.

In the present investigation 27 licensed physicians, all holding the medical doctorate, served as judges. They judged results (levels) from nine frequently utilized numerical medical information indices which were varied systematically and independently. An upper and lower limit for each of the nine indices was determined from the medical literature and on the advice of medical consultants. Specific stimuli within these limits were spaced in equal logarithmic steps when feasible. Judgments were made relative to "degree of concern" for a contrived 35 year old patient's health status. The clinical laboratory indices and the various levels of each index were presented to the judges in randomized orders.

For pulse, red blood count, white blood count, and specific grabity of urine, levels above and below normal were presented. It was believed that the most appropriate subjective "concern" scales could be developed when the levels above and below normal were considered as separate scale-indices. The same analyses were conducted with both the above and below normal levels. With this in mind, a total of 13 psychophysical scales were developed.

A high degree of inter-judge concordance or agreement was found for each of the 13 scales. This is notable since physicians from six different medical specialities were included in the sample. This is even more remarkable as several judges could not recall having been exposed to any formal didactic experience in which degree of concern, suspicion, or similar concepts had been linked to specific levels on any of the stimulus continua. A rather thorough review of medical literature by the investigator also did not reveal any pairings of such concepts or percepts.

In general the results indicate that for nine of these medical indices the relationships observed were curvilinear when degree of concern was plotted against the appropriate stimulus metric. Log-log transformations rectified the data reasonably well. This would imply that a power function model could provide an appropriate description for these data. With four of the indices (protein, sugar, systolic, and diastolic blood pressure) the relationships, when degree of concern was plotted against the stimuli, were markedly linear in nature. It was suggested that the underlying judgmental continua for these four indices are perhaps metathetic. Physicians may view these information indices as some sort of ordered category measures even though the underlying laboratory measures are continuous in nature.

Inter-scale comparisons were also made from estimations suggested by the judges. It was suggested that upper levels on temperature, red blood count, and both blood pressure indices appeared to be subjectively associated with higher degrees of concern. By contrast, all utilized levels on the specific gravity of urine index failed to elicit much subjective concern in contrast to the other scales. Possible reasons for these were discussed. The inter-scale comparisons would seem to hold potential value for future investigations with regard to selection of possible standards and assignment of numerical modulus values.

For specific gravity, pulse, and red blood count it was found that the utilized standard and its prescribed modulus was disparate from that used intrinsically by the judges. The net effect of this disparateness was to increase judgmental variability or "noise" in the measurement system.

Implications for the direct estimation literature seem clear. The present investigation represents one of the earliest successful extensions of these methodologies into the topic area of clinical judgmental processes. It was suggested that direct estimation procedures are sufficiently sensitive to assist in the clarification of the many enigmatic ambiguities now existent in the clinical judgment literature.

Possible implications for medical education were also drawn. The development of scales similar to those produced in the present inquiry could provide valuable communication vehicles whereby the exigencies of the physician's office could be brought into the classroom.

Limitations of the present investigation were also discussed. The sample of physician-judges was in no way random, thereby restricting possible generalizations. Methodological considerations such as appropriateness of standard numerical modulus values, and possible scale interactions were also discussed. Suggestions for future research into this general area of medical judgment were also made.

APPENDIX

Temperature

Stimulus Level	Concern Scale Value*	Standard Deviation
98.6	0.037	0.036
99.0	0.867	0.756
101.2 (Standard Stimulus)	1.699	-----
102.6	1.828	0.378
104.0	2.114	0.203
105.4	2.328	0.321
108.2	2.593	0.300
Pulse		
68	0.794	0.834
82 (Standard Stimulus)	1.699	-----
98	1.257	0.758
108	1.661	0.610
141	2.095	0.191
168	2.215	0.210
201	2.368	0.270
Sugar		
trace	1.018	0.765
1+ (Standard Stimulus)	1.699	-----
2+	1.952	0.203
3+	2.137	0.166
4+	2.224	0.211
Protein		
trace	1.025	0.721
1+ (Standard Stimulus)	1.699	-----
2+	1.959	0.213
3+	2.119	0.241
4+	2.291	0.265
Specific Gravity		
1.000	2.015	0.460
1.003	1.633	0.713
1.010	1.344	0.726
1.013 (Standard Stimulus)	1.699	-----
1.016	0.969	0.841
1.020	1.021	0.848
1.026	1.214	0.804
1.030	1.410	0.781
1.040	1.574	0.889

* Values for scale values and standard deviations are in logarithmic units.

Diastolic Blood Pressure

Stimulus Level		Concern Scale Value*	Standard Deviation
80		0.300	0.627
93	(Standard Stimulus)	1.699	-----
108		1.976	0.202
125		2.123	0.193
145		2.253	0.236
168		2.320	0.244
195		2.431	0.284
220		2.511	0.311
		Systolic Blood Pressure	
108		0.681	0.843
130	(Standard Stimulus)	1.699	-----
143		1.986	0.829
173		1.986	0.213
191		2.108	0.217
211		2.215	0.219
232		2.312	0.201
252		2.372	0.247
281		2.439	0.282
		Red Blood Count	
1.0		2.415	0.346
2.5		2.108	0.508
4.0		1.115	0.939
5.5		0.689	0.848
6.0		0.855	0.829
7.0	(Standard Stimulus)	1.699	-----
8.5		1.571	0.674
10.0		1.899	0.505
13.0		2.172	0.532
		White Blood Count	
750		2.489	0.309
2,250		2.113	0.203
5,250		0.626	0.830
7,500		0.448	0.709
9,500		0.803	0.846
10,500	(Standard Stimulus)	1.699	-----
24,750		2.114	0.241
100,000		2.373	0.361
200,000		2.571	0.365

* Values for scale values and standard deviations are in logarithmic units.

REFERENCES

Altman, P. L. and Ditmer, D. S. (Eds.) Biology data book. Washington, D.C.: Fed. of American Societies for Experimental Biology, 1964.

Block, W. E. A conceptual framework for the clinical test situation. Psychological Bulletin, 1964, 61, 168-175.

Boring, E. G. A history of experimental psychology. New York: Appleton-Century-Crofts, 1957.

Cullen, M. F. The multitest laboratory in health care of the future. Hospitals, 1964, 41, 1-5.

Eisler, H. Empirical test of a model relating magnitude and category scales. Scandinavian Journal of Psychology, 1962, 3, 88-91.

Ekman, G. and Künnapas, T. A further study of direct and indirect scaling methods. Scandinavian Journal of Psychology, 1963, 4, 77-80.

Ekman, G. and Sjöberg, L. Scaling. Annual Review of Psychology, 1965, 16, 45

Erdman, A. J., Jr. Data processing and the interpretation of symptoms, I. Results. In J. A. Jacques (Ed.), The diagnostic process. Ann Arbor, Mich.; Malloy Litho. Inc., 1964.

Fillenbaun, S. Contextual effects in judgment as a function of restrictions in response-language. American Journal of Psychology, 1963, 76, 103-109.

Goldberg, L. R. Simple modes or simple judgments? American Psychologist, 1968, 23, 483-495.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Harder, T. J. and Gow, A. E. The essentials of medical diagnosis: A manual for students and practicioners. (2nd ed.) New York: Williams and Wilkens, 1953.

Hermann, J. and Fox, R. Magnitude estimation of sexual standards. Perceptual and Motor Skills, 1967, 24, 88.

Hunt, W. A. An actuarial approach to clinical judgment. In B. Buss and I. A. Berg (Eds.), Objective approaches to personality assessment. Princeton, N. J., Van Nostrand, 1959, Pp. 51-73.

Hunt, W. A. and Jones, N. F. The experimental investigation of clinical judgment. In A. J. Bachrach (Ed.), Experimental foundations of clinical psychology. New York: Basic Books, 1962, Pp. 26-31.

Hunt, W. A. and Schwarz, M. L. Reliability of clinical judgments as a function of range of pathology. Journal of Abnormal Psychology, 1965, 70, 32-33.

Indow, T. An example of motivation research applied to product design, published in Japanese in Chosa to Gifutsu No. 102 (1961). Cited by S. S. Stevens. A metric for social consensus. Science, 1966, 151, 530-541.

Koh, S. D. Scaling of musical preferences by the mentally retarded. Science, 1966, 151, 432-433.

Ledley, R. S. and Lustead, L. B. Reasoning foundation of medical diagnosis. Science, 1959, 130, 9-15.

Lewis, D. Quantitative methods in psychology. New York: McGraw-Hill, 1960.

Lipkin, M. The role of data processing in the diagnostic process. In J. A. Jacquez (Ed.), The diagnostic process. Ann Arbor, Mich: Malloy Litho. Co., 1964, Pp. 315-346.

Meehl, P. E. Clinical versus statistical prediction. Minneapolis: University of Minnesota Press, 1954.

Miller, S. E. (Ed.) A testbook of clinical pathology (5th ed.), Baltimore: Williams and Wilins, 1955.

Perole, S. J. The relation between category, rating and magnitude estimation judgments of occupational prestige. American Journal of Psychology, 1963, 76, 395-403.

Poulton, E. C. The new psychophysics: six models for magnitude estimation. Psychological Bulletin, 1968, 68, 1-19.

Sellin, T. and Wolfgang, M. E. The measurement of delinquency. New York: Wiley, 1964.

Sodeman, W. A. and Sodeman, W. A., Jr. Pathologic physiology: Mechanisms of disease, (4th ed.), Philadelphia: Saunders, 1967.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.) Handbook of Experimental Psychology. New York: Wiley, 1951, Pp. 1-49.

Stevens, S. S. On the psychophysical law. Psychological Review, 1957, 153-179.

Stevens, S. S. and Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. Journal of Experimental Psychology, 1957, 54, 377-411.

Stevens, S. S. Problems and methods of psychophysics. Psychological Bulletin, 1958, 55, 177-195.

Stevens, S. S. Review of L. L. Thurstone: The measurement of values. Contemporary Psychology, 1959, 4, 388-389.

Stevens, S. S. A metric for the social consensus. Science, 1966, 151, 530-541. (a)

Stevens, S. S. On the operation known as judgment. American Scientist, 1966, 54, 385-401. (b)

Stone, L. A. Psychiatrists' prognostic judgments regarding functional psychotic disorders: A prognostic scale. Behavioral Science, 1966, 11, 115-120.

Stone, L. A. Clinical psychophysics. Studia Psychologica (Bratislava), 1968, 10, 161-173. (a)

Stone, L. A. Judged prognosis: An attempt to apply a psychophysical approach to scale psychiatric opinion. Journal of Clinical Psychology, 1968, 24, 28-32. (b)

Stone, L. A. Bases for psychiatric impairment severity judgments: Psychophysical power functions? Behavioral Science, 1969, 14, 133-137.

Stone, L. A. and Skurdal, M. A. Judged prognosis for functional psychosis disorders classifications: A prothetic continuum? Journal of Consulting and Clinical Psychology, 1968, 32, 469-472.

Torgerson, W. S. Quantitative judgments scales. In H. Gulliksen and S. Messick (Ed.) Psychological scaling: Theory and applications. New York: Wiley, 1960, Pp. 21-31.

Underwood, B. J. _Psychological research_. New York: Appleton-Century-Crofts, 1957.

Warren, R. M. and Warren, R. P. A critique of S. S. Stevens' "New Psychophysics." _Perceptual and Motor Skills_, 1963, _16_, 797-810.