$2.25

Research Series No. 28

THE CONSISTENCY OF READING DIAGNOSIS

John F. Vinsonhaler

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

Printed and Distributed
by the
College of Education
Michigan State University

June 1979

This work is sponsored in part by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-76-0073)

Institute for Research on Teaching

The Institute for Research on Teaching was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Abstract

Three types of clinical agreement are examined as they pertain to collection and diagnosis: (1) group agreement, (2) intra-clinician agreement, and (3) inter-clinician agreement. Results indicate that in diagnosing reading cases, reading clinicians do not agree very well with themselves or with other clinicians. Results are discussed in the light of the Inquiry Theory.

The Consistency of Reading Diagnosis[1]

John F. Vinsonhaler[2]

Introduction

Studies of clinical problem solving have been a part of the psychological literature for several decades. Although these studies have taken diverse forms, a common thread of intent links them: the desire to understand the psychological processes by which clinicians arrive at diagnoses and plans for therapy. Historically, the major work on clinical problem solving has been in medicine. These studies have included investigations of judgment (Feinstein, 1967), decision making (Schwartz, Gorry, Kassirer, & Essig, 1973), problem solving and computer-aided diagnosis (DeDombal, Leaper, Staniland, McCann, & Horrocks, 1972). Given the long history of research in medicine it was inevitable that investigations of clinical problem solving would eventually be extended into other fields, including education, and behavioral science.

In 1966, DeDombal and his colleagues introduced the results of painstaking studies of surgical decision making into the training of medical students. The results were spectacular (DeDombal, Horrocks, Clamp, & Storr, 1974). Earlier, studies of psychological and psychiatric

---

[1]This paper was presented at the 1979 International Reading Association meeting in Atlanta, Georgia under the title: "Clinical problem solving among reading specialists: The problem of agreement on a diagnosis."

[2]John Vinsonhaler is an MSU professor of educational psychology and coordinator of IRT's Clinical Studies group. The group's other members are George Sherman, Linda Patriarca, Christopher Wagner, Ruth Polin, Doron Gil, Joel VanRoekel, and Annette Weinshank. Jay Stratoudakis, Ethelyn Hoffmeyer, and Roger Frame are also associated with the group as researchers supported by non-IRT funding.

problem solving were performed and implications clearly drawn for education. In about the middle 1970s, a number of articles appeared on clinical problem solving in education. These reports included studies on training for "trouble shooting" (diagnosis in electronic and computer technology), learning disabilities (Lerner & Schuyler, Note 1; (Vinsonhaler, Wagner, & Elstein, 1977).

IRT research in reading and learning disabilities has its roots in the investigation of clinical problem solving in internal medicine (Elstein, Shulman, & Sprafka, 1978 ; Barrows, Feightner, Neufield, & Norman, Note 2, and Vinsonhaler, Wagner, & Elstein, 1977). From these "inquiry studies" in medicine, we in the Clinical Studies Program at the Institute for Research on Teaching drew three principles to direct our studies of reading clinicians: (1) the practical value of a well understood theoretic base for empirical research on clinical problem solving; (2) the efficacy of methods developed in medicine to examine problem solving under well controlled conditions using simulated cases and stimulated recall interviews; and (3) the need for a systematic program of research studies which share a common methodology.

This paper summarizes our attempts to apply these principles to the study of reading diagnosis and remediation. The first section discusses the theoretic basis of our research. The second section describes application of the medically-oriented research methods in reading and the major results. Finally, the last sections indicate how this initial study has necessarily led to a systematic program of research studies in the Institute for Research on Teaching.

Theoretic Basis of this Study

Agreement between clinical judgments is a classic problem in medicine. For many years studies have periodically appeared investigating the agreement of physicians' medical judgments. While some of these studies have shown substantial agreement among physicians, others have documented marked disagreement (Garland, 1959). Examples of the latter include:

1. The diagnosing of myocardial infarction from ECG (56% error rate based upon autopsy results reported by Paton, 1957);

2. The diagnosing of pulmonary disease from x-ray photographs, where disagreement of the physician on the diagnosis was generally 20% with himself and 30% with other radiologists (Fletcher, 1952; Cochrane and Garland, 1952; Yerushalmy, 1955, 1969); and

3. The diagnosing of various psychiatric disorders (Kendell, 1975).

Given these findings about agreement among physicians, we decided that an investigation of agreement among reading clinicians might also yield significant results.

The theoretic basis of our studies is the Inquiry Theory (Vinsonhaler, Wagner, & Elstein, Note 3); it is a corollary of that theory which shaped this study.[2] We examined the implications of the Inquiry Theory for agreement on diagnosis and remediation. (In the present paper we shall limit our discussion to diagnosis. Remediation will be reported later.) One assumption of the Inquiry Theory is that diagnostic decisions are probabilistically determined by the case, clinical memory, and clinical strategy. From this assumption we derived the following corollary by means of computer simulation studies and set theoretic arguments:

Agreement Corollary:

For a given case and a given set of clinicians, the greater the similarity of clinical memory, the greater the agreement of diagnoses. Thus, assuming a memory (denoted Mi for the i^{th}

[2]See Appendix A for a brief description of this theory.

clinician) which is composed of discrete elements (i.e., representations of cues, diagnostic categories, relations between cues and categories, and strategies):

If $N(Mi \cap Mj) \geq N(Mk \cap Ml)$, then $N(Dxi \cap Dxj) \geq N(Dxk \cap Dxl)$.

Where $N(Mi \cap Mj)$ denotes the number of common elements in two memories and $N(Dxi \cap Dxj)$ denotes the number of common elements in two diagnoses.

Three particulars of the Agreement Corollary were investigated in the present study.

1. *Group Agreement:* Measures of agreement involving the comparison of individual diagnoses with group diagnoses should be greater than or equal to comparisons involving only individuals.

2. *Intra-Clinician Agreement:* Measures of agreement of an individual's diagnosis with his/her diagnosis on equivalent cases should be greater than measures of agreement of the individual's diagnosis with those of other clinicians.

3. *Inter-Clinician Agreement:* Measures of agreement between clinicians for the same case should reflect the similarities of the memories elicited by the cases at hand. Low agreements would suggest a lack of common memories, strategies, and procedures among the group of clinicians.

Since the same arguments might also be applied to agreements among clinicians on what cues to collect, all of the above implications seem to hold for cue collection.

In summary, the major emphasis of the present study was upon the agreement corollary as applied to diagnoses and the cue collection procedure.

Method

The purpose of this initial observational study in reading was threefold: first, to determine the feasibility of using the Inquiry Theory methods for the study of clinical problem solving in reading; second, to obtain empirical data on the agreement of reading clinicians on diagnosis and remediation; and third, to examine the implications of the results for the Shulman/Elstein (Elstein, Shulman, & Sprafka, 1978) theory. Our general intent was to study reading clinicans' problem solving under the most supportive possible conditions.

Subjects

The subjects were not randomly sampled from a population. Rather, every attempt was made to recruit the most senior and most respected practicing clinicians in the mid-Michigan area. Candidates were recommended by university faculty and/or school administrators. From those nominated, a set of eight volunteer subjects were selected on the basis of multiple recommendations. All subjects had master of arts degrees in reading and had been practicing clinicians or reading specialists for at least five years. All were experienced reading teachers, and most had, or were in the process of obtaining a Ph.D. in reading. Not all clinicians had been trained in Michigan. Some had received their training at eastern and other midwestern universities. Half of the subjects were currently reading faculty members at Michigan State University. All subjects were paid at professional rates for their participation.

Design of the Study

Simulated cases of reading problems were used in this study. Each case was based upon a real child who was once a client of the Reading Clinic at Michigan State University. The reading problems represented by the cases included sight word deficiencies, inadequate structural and phonetic analysis skills, inadequate fluency of oral reading, and poor comprehension. The cases have been described by many clinicians as representative of the problems most frequently encountered in the public schools. A total of four clients were used in generating the eight simulated cases. Thus each simulated case had two equivalent forms. Equivalent forms were prepared by making minor changes in the original data bases. For each case, an inventory was provided describing

the data available (that is, the cues that could be collected concerning the case's reading problems).

All stimulus materials (including simulated cases, equivalent forms, and randomly ordered data inventories) were subjected to counter balancing to minimize systematic effects. In general, procedures were designed to provide optimum support for the subjects. No limit was placed upon time or the number of cues which could be collected during the observational session. An inventory of available data was used to help the clinician select the information needed. Subjects were requested to prepare their diagnostic and remedial reports as they normally would, but no time or length restrictions were placed upon them.

The same four-step procedure was used for each simulated case run. First, the subjects were given instructions and practice with a sample case. Second, there was an observational session, directed by an experimenter and recorded by a clinical observer, in which the subjects collected data using an inventory of available data. Third, the subjects prepared a written diagnostic and remedial report. Fourth, there was a debriefing session, directed by a clinical observer aided by an experimenter, in which the subjects underwent stimulated recall. Each data item was presented along with controlled interview questions: "Why did you request this? What did it tell you?" Each clinician received three simulated cases at one week intervals. The first and third cases were equivalent forms of the same case.

Data Analysis

The Clinical Studies group has worked on methods of measuring agreement in diagnostic and remedial decision making for more than two years. At present, we seem to obtain the most methodologically sound results with the basic methods described below. All of these methods require that the raw set of diagnostic statements (or cue requests) be translated into a standard vocabulary for diagnosis or cue description.

The standard vocabulary translation of the diagnoses proved to be much more difficult in reading than in medicine. The major problem was the lack of a generally agreed upon vocabulary for diagnosis and remediation. The method of translation used in the present study was to first allow clinicians to use natural language in specifying their diagnoses; then, these diagnostic statements were sorted by staff clinicians into groups, and each group was given a standardized label.

A reliability check was made by having the clinicians sort a set of randomly selected statements a second time (10% of the original set were sorted three months after the first categorizations). Some 75% of the statements were identically sorted the second time. When highly non-specific categories (e.g., "general statements about phonics," "general statements about health," etc.) were dropped, agreement increased to 81% on the second sorting. A total of 160 categories or groups of statements were found to characterize the set of simulated cases, although each case involved no more than 50 diagnostic categories.

Many attempts have been made over the past two years to reduce the number of categories, since most agreement measures are sensitive to vocabulary size. The major problem is that combining categories yields large numbers of logical inconsistencies (e.g., the same diagnosis often includes inconsistent statements such as "no problem with phonics" and "needs work on long vowels"). Given this caution about potential methodological difficulties, I will describe our measurement of agreement.

The *proportional agreement* statistic is a measure of group agreement. First, a "domain of statements" is defined (i.e., a set of standard statements for diagnosis or cue requests). Second, the proportion of

diagnoses, clinicians, or encounters including each statement is calculated. This statistic gives an overall indication of the diagnostic and cue categories most frequently employed by clinicians.

The *commonality score* is a measure of individual-to-group agreement developed in medicine and subsequently modified for use in education. Basically, the score reflects the agreement of any individual diagnosis or cue selection with that of a group. The statistic is always between zero and one; it uses only part of the information available on agreement (i.e., only the proportional agreement statistic is used in calculating an individual's score).

The *inter-clinical correlation* is a measure of the agreement of one clinician with another clinician in the cue requests or diagnostic statements elicited by the same simulated case. The method requires (1) the definition of a domain of statements; (2) the categorization of statements as being present in or absent from an encounter (e.g., included in or excluded from a diagnosis). A two by two contingency table is prepared and a PHI coefficient calculated. Cells in the table include the number of statements: (1) present in both encounters; (2) present in the ith but not the jth encounter; (3) present in the jth but not the ith encounter; and (4) statements not present in either encounter. Thus, this statistic uses the full set of agreement data.

The *intra-clinician correlation* measures the agreement of clinicians with themselves on encounters with equivalent forms of the same simulated case. This coefficient is calculated using the same PHI coefficient/ two by two contingency table procedure described for the inter-clinician correlation, but applying it to two diagnoses by the same clinician.

Many methodological problems have plagued us in the development of these dependent variables. I have already discussed the difficulties with the standardized vocabulary analysis. Another problem was the commonality score, which seems to yield artificially high levels of agreement. Similarly, the PHI coefficients, when corrected for unequal marginal frequencies, definitely yielded artificially low (even negative) correlations.[3] Suffice it to say that we and our colleagues in the Institute view these methods of agreement measurement as the best presently available,[4] but caution the reader that the methods may include methodological difficulties beyond those currently recognized.

Results

In the introduction I acknowledged our hypothesis that the Inquiry Theory methods using simulated cases with stimulated recall might yield significant insights into the clinical problem solving of reading clinicians. In the following sections I will report results on this hypothesis with respect to three general areas: (1) diagnostic agreement (the degree to which clinicians agree on the diagnosis of cases of reading difficulty), (2) cue collection agreement (the degree to which clinicians agree on the particular data that should be collected for given cases of reading difficulty), and (3) informal observations on methods of diagnosis.

Agreement in Diagnosis

Figure 1 shows results for the proportional agreement statistic for

---

[3]The recommended correlation is PHI (corrected) = PHI/PHI (Maximum; where PHI (maximum) is the largest possible PHI coefficient given existing marginal frequencies of the contingency table (Cureton, 1959; for a critique, see Carroll, 1961).

[4]Many methods of estimating the association between diagnoses have been tried, e.g., Kappa (Cohen, 1969). All seem to yield similar results as uncorrected PHI.

Figure 1: Diagnostic statements mentioned most frequently*

Diagnostic Statement From Standard Vocabulary	Proportion of Diagnoses Including Statement By Simulated Case			
	Case S	Case M	Case D	Case T
At least average reading potential	.67	.33	.50	.67
Adequate verbal skills	.33	.50	.50	
Poor oral reading	.50	.67	.33	
Problems with vowels	.50	.33	.33	
Sight words low	.33		.83	
Phonics weak	.33			.67
Auditory acuity problem		.50		.67
Consonant blends not a problem	.33		.33	
Good use of context	.33		.33	
Writing problem	.33	.33		
Spelling problem		.33		.33
Normal interest and behavior			.33	.33
Attitude toward reading poor	.50			
No problem with isolated letter sound skills	.50			
Speech problem		.50		
Problem with syllables		.50		
Handwriting problem		.50		
Problem with visual memory			.50	
Health problems in school				.50
Poor word analysis skills				.50
Auditory discrimination problem				.50

*Statements mentioned in 50% of the diagnoses for a single simulated case, or in the diagnoses for 50% of the cases, or both.

the most commonly-mentioned diagnostic statements. Proportional agreement for a given diagnostic statement is the proportion of diagnoses mentioning that statement. In general, the statements in the figure seem to show a reasonable level of group agreement and largely reflect what we judge to be correct statements about the cases. For example, the most frequently mentioned category -- average reading potential -- is correct for all of our simulated cases. Many diagnostic categories are shared by cases, but some are used uniquely for given cases.

The set of statements seems to differ markedly from the type given in medical diagnosis in several respects. First, the reading diagnoses are much longer and include many more categories. Second, the reading diagnoses include many observations and statements of strengths as well as weaknesses, whereas medical diagnoses tend to be limited mainly to problem statements. Third, the categories used reflect relatively few "causal" factors for reading problems, whereas medical diagnoses usually include mainly causal statements. In summary, the set of categories obtained from our reading diagnoses could be best described as what might be produced by a group of physicians performing a routine health examination.

Figure 2 summarizes our results for agreements between diagnoses. The first four columns are results for individual cases. The last two columns summarize results of our study of reading and a comparable study of medical practitioners. The medical study is based upon "differential diagnoses", i.e., the set of probable diagnoses assembled prior to the use of specific laboratory tests for a final medical diagnosis. Since highly precise laboratory findings are not generally available in reading, the medical differential diagnosis seems more appropriate for comparison with reading. The PHI coefficients reported are not corrected for marginal inequalities.

The major findings are most obvious in the right-most columns of the figure. First, the commonality score (developed in medicine) shows little difference from reading to medicine. Second, the intra-diagnosis correlations are quite low for reading. Only an average of 1% of the variance in the diagnosis of the second presentation of a case can be predicted from the diagnosis of the first presentation of the same (equivalent form) case. Third, the intra/inter diagnoses correlations are markedly lower in reading than in medicine -- even though the medical diagnoses are based only upon history taking and physical examination results. Finally, note that the prediction (that intra-clinician correlation should exceed inter-clinician correlation)seems to be borne out by the data (mean intra-PHI is +.13, while intra/inter-PHI is -.07).

I would emphasize that these results are based upon a very small sample of subjects and are clearly open to questions on methodological grounds, as noted previously. For these reasons, we have not used inferential tests. For such small sample studies we prefer to depend upon replications of gross effects rather than inferential statistics.

Agreement in Cue Collection

In general, the results for cue collection show a substantially higher level of agreement than do the diagnostic results. For the four cases, the mean commonality score was .74 with a standard deviation of .14. The mean intra-clinician correlation (PHI) was .33 with a standard deviation of .27. Intra/inter correlation for cues collected yielded a mean of .15 and a standard deviation of .23.

In summary, reading clinicians seem to show a higher level of agreement with themselves and others in the data collected during the

Figure 2: Agreement statistics for diagnoses of four cases in reading and medicine*

Agreement	Case S	Case M	Case D	Case T	Reading Clinicians	General/Internal Medicine
Number of clinicians	3	3	3	3	8	37
Number of diagnoses	6	6	6	6	24	58
Commonality Score						
Mean	.53	.52	.54	.59	.55	.72
Standard Deviation	.14	.23	.19	.17	.18	.21
Intra Diagnosis Correlation (PHI)						
Mean	.09	.27	.04	.12	.13	-
Standard Deviation	.13	.16	.17	.01	.14	-
Intra/Inter Diagnosis Correlation(PHI)						
Mean	-.10	-.09	-.04	-.05	-.07	.34
Standard Deviation	.16	.22	.16	.15	.17	.28

*Based on data provided by Barrows et al,(Note 2). Agreement is for differential diagnosis (based only on symptoms and signs -- no laboratory data) of four common problems in general medicine. Statistic reading diagnoses are by case average over four cases.

clinical encounter than they do in stating the diagnosis based upon such cue collection. Further, there are clear clinician preferences for some types of information, such as background data on school behavior and measures of reading potential and reading skills.

Some Conjectures on Methods

The study of the process by which our clinicians made their decisions is still under analysis. However, a few observations may be worth comment.

First, two types of strategy seem to characterize clinicians: deductive and inductive problem solving. Deductive problem solvers tended to direct the inquiry process by the use of hypotheses about the client's reading. Cues were collected to test specific hypotheses. Inductive problem solvers, on the other hand, tended to use cue collection to direct the inquiry. Thus, certain types of cues were collected, then statements (including diagnostic statements) were generated from the resulting data base. While both approaches are observable in medicine, deductive thinking predominates. In reading, the reverse seems true. Most reading clinicians used inductive strategies.

A second observation is that the strategy and memory used by our clinicians seemed to be based upon (1) a model or theory of the process of reading and learning to read and (2) clinical experience under conditions of feedback (e.g., reaction to remediation). Thus, many clinicians voiced some overall conception of reading problems, sometimes very general (e.g., "reading problems are mainly motivational problems") and sometimes quite precise (e.g., "reading problems are mainly the failure to learn skills -- sight word, word analysis, fluency, and comprehension"). Most clinicians, however, tempered these models with references to specific cases from clinical experience where remediations had been successful.

Discussion

In the introduction to this paper I noted three principles, drawn from the Inquiry Theory research in medicine, which have guided our work in reading. The results of the present study will be discussed in relation to these principles, and then tentative conclusions summarized.

The first principle taken from prior medical research argues for the practical value of the inquiry theoretic base for research on clinical problem solving. It would seem to us that the Inquiry Theory has frequently led us away from errors in method and directed our work into potentially useful channels. Two examples come quickly to mind. The agreement corollary predicted positive, zero (or at worst, small negative) correlations among diagnoses. In two of our experiments, widely used PHI coefficient corrections yielded high negative values. Careful study prompted by our theoretic expectations uncovered an artifact: conditions yielding small negative correlations also yielded maximal corrections for unequal marginal frequencies (e.g., correlations of -.10 could be artificially set to -1.00). The second example concerns the commonality statistic. In medicine, the commonality score had been accepted as a valid measure of agreement between individuals via the use of a group consensus, and, indeed, in medicine the statistic may be valid for this purpose. The Agreement Corollary of the Inquiry Theory suggested the possibility of differences in individual and group agreement, thus encouraging the use of additional statistics reflecting the agreement between individual diagnoses; e.g., correlations and proportions of agreement. Such statistics yielded substantially different results from the commonality scores. This result, we believe, reflects the relatively low reliability of individual versus group

diagnosis in reading. If established as empirically valid, this low individual reliability has profound implications for the training and practice of reading specialists.

The second principle taken from the medical research was that the inquiry methods are of value in studying clinical problem solving in reading. By "of value," we mean contributing to the understanding of the clinical encounter. Consider the empirical findings we have uncovered.

First, we have developed methods differentiating individual from group agreement. This differentiation is important, because clinical decisions are largely made by individuals and, in both reading and medicine, individual agreement has been shown to be far lower than group agreement measures. (In fact, some individual agreements are zero in both reading and medicine.)

Second, the medical phenomenon of deductive versus inductive reasoning strategies has also been observed in reading, and the nature of the phenomenon more clearly understood. Thus, either strategy may be emphasized by the same clinician, but one or the other is more typical of a given clinician. Further, early hypothesis generation appears to be a logical consequence of deductive reasoning (Vinsonhaler, Wagner, & Gil, Note 4).

Finally, the reading studies have generated a number of promising theoretical concepts and have increased our understanding of clinical decision making. For example, the "model of process" hypothesis poses that the major source of clinical memory and strategy is a clinician's theory about the processes underlying the phenomena s/he is attempting to influence (e.g., a theory of reading or the physiology of immunological reactions). Clinical memories and strategies are initially deduced from

model of process, but become directly associated with cue patterns given repeated diagnostic practice under conditions of feedback (i.e., reaction to treatment based on a diagnosis). Another example, the "standard vocabulary" hypothesis, states that the agreement corollary is valid only where clinicians share a common technical vocabulary of reasonable precision.

The third principle taken from medical studies asserts the value of a systematic program of research studies which share a common theoretic and methodological base and a set of interrelated purposes. We are presently testing this notion with a series of studies designed to determine the probable causes of the low intra/inter-clinician agreement observed in this study. (These are summarized in Appendix B). Each is aimed at the evaluation of a different explanation of our results.

Conclusions

The conclusions offered here are tentative, and are based not only upon the findings of this study, but also upon preliminary results of the studies described in Appendix B.

The Inquiry Theory and methods adapted from medical research seem to have worked very well in our studies of reading. The theory has been a good guide, the methods have yielded promising theoretic improvements, and both have led to an interesting, coherent plan for further research.

The major empirical finding of our study seems to be that *in diagnosing cases, reading clinicians do not agree very well with themselves or with other clinicians,* regardless of what agreement statistic is used. I have reported here the PHI coefficient, which showed essentially very low or zero agreement. Other statistics, such as Chi square, contingency

coefficients, and various proportional agreement statistics have all yielded similar results. I have cautioned the reader about the tentativeness of such findings, and the many possible artifactual complications of our statistics. Still, similar results are rapidly accumulating from more recent studies.

If this low agreement proves the rule, we should not castigate reading clinicians, since similar results have been uncovered in many specialized fields of medicine and psychology. Instead, we must seek a better understanding of the causes of low diagnostic agreement and methods by which we may use this knowledge to improve the training and decision making of clinicians. This task is the primary goal of our research.

References

Carroll, J.B. The nature of the data, or how to choose a correlation coefficient. Psychometrica, 1961, 26, 347-371.

Cochrane, A.L., & Garland, L.H. Observer error in the interpretation of chest films: International investigation, Lancet, 1952, 2, 505-509.

Cohen, J.A. Coefficient of agreement for nominal scales. Educ. Psychol. Measmt. 1969, 20, 37-46.

Cureton, E.E. Note on PHI/PHI (MAX). Psychometrica, 1959, 24, 89-91.

DeDombal, F.T., Horrocks, J.C., Clamp, S.E., & Storr, J.E., Simulation techniques and computer-aided teaching of the clinical diagnostic process: Five years experience. Medinfo '74 Proceedings. New York: North Holland Publishing Co., 1974. pp. 247-252.

DeDombal, F.T., Leaper, D.J., Staniland, J.R., McCann, A.P., & Horrocks, J.C. Computer-aided diagnosis of acute abdominal pain. British Medical Journal, 1972, 2, 9-13.

Elstein, A., Shulman, L.S., & Sprafka, S. Medical problem solving: An analysis of clinical reasoning. Cambridge, MA.: Harvard University Press, 1978.

Feinstein, A.R. Clinical judgment. Baltimore, MD.: Williams and Wilkins, 1967.

Fletcher, C.M. Clinical diagnosis of pulmonary emphysema: Experimental study. Proceedings of the Royal Society of Medicine, 1952, 45, 577-584.

Garland, L.H. Studies on the accuracy of diagnostic procedures. American Journal of Roentgenology, 1959, 82, 25-38.

Kendell, R.E. The role of diagnosis in psychiatry. Oxford, Great Britain: Blackwell Scientific Publications, 1975.

Paton, B.C. The accuracy of diagnosis of myocardial infarction. A clino-copathologic study. American Journal of Medicine, 1957, 23, 761-768.

Schwartz, W.A., Gorry, G.A., Kassirer, J.B., & Essig, A. Decision analysis and clinical judgment. American Journal of Medicine. 1973, 55, 459-472.

Yerushalmy, J. Reliability of chest radiography in the diagnosis of pulmonary lesions, American Journal of Surgery, 1955, 89, 231-240.

Yerushalmy, J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. Radiologic clinics of North America.1969, 7, 381-391.

Appendix A

The Inquiry Theory

The theoretic basis of our work is the "Inquiry Theory" developed by L.S. Shulman and A.S. Elstein and documented in their book on medical problem solving (Elstein, Shulman, & Sprafka, 1978). In 1973, Shulman, Elstein, and colleagues began work on a more formal version of the Inquiry Theory involving the use of computers to simulate medical clinicians who solved problems in exact accord with the theory (Vinsonhaler, Wagner, & Elstein, Note 2). Such formal statements of the Inquiry Theory have permitted the explication of the theory as a set of assumptions and a set of principles, termed corollaries, deducible from the assumptions. Since complete statements of the Inquiry Theory are available (Vinsonhaler, Wagner, & Elstein, Note 2), I shall merely describe the theoretic structure.

The Clinical Encounter

The first principle of the Inquiry Theory states that the behavioral domain addressed by the theory involves a clinician, a case or patient, and an interaction yielding a decision on the diagnosis (the state of the case) and the therapy (how this state can be improved). This assumption, of course, places major limitations on the theory (e.g., see Elstein, et al, 1978); but the assumption also permits a reasonably well defined behavioral domain for study.

The Simulated Case Assumption

The second principle of our theory is that important problem-solving behaviors of clinicians can be elicited through simulated cases or patients.

We fully recognize that not all the behaviors necessary for being an effective clinician are elicited by a simulated case, but it is our assumption that at least some of the important cognitive skills are elicited by sets of data gathered about real cases with real problems and stored for presentation as simulated cases. This assumption seems reasonably valid in medicine based upon the wide use of simulated cases in training and in research on validity. Regretfully, legal limitations on the use of subjects and lack of funding have prevented us from obtaining direct data on the validity of this assumption in reading.

The Simulated Clinician Assumption

The third principle of our theory is that clinical problem-solving behavior is probabilistically determined by clinical memory, clinical strategy, and the nature of the case. In diagnosis, for example, clinical memory is assumed to contain representations of: (1) cues (observable findings or combinations of findings); (2) diagnostic categories (e.g., problems like myocardial infarction or poor oral reading) and (3) relationships between problems and cues permitting the diagnosis (the inferences of a set of problems from a particular set of cues). Similar memory structures are assumed for therapeutic decision making. This simulated clinician assumption is supported by numerous research findings (e.g., Elstein et al., 1978).

Appendix B

The IRT Clinical Studies Research Agenda

Possible Explanations of low Dx/Cx Agreement	Existing or planned research
Non-replicable "chance" finding, unique to reading clinicians	Observational Study of Reading and Learning Disability clinicians. Twenty clinician's performance on two simulated cases (one reading problems; one LD and reading problems)
Non-replicable "chance" finding, unique to clinicians	Observational Study of classroom teachers. Ten teachers performance on two classroom oriented simulated cases.
Non-replicable "chance" finding; due to lack of standardized vocabulary among clinicians	Observational Study replication of one reported in this paper, using standardized checklists permitting subjects to transcribe diagnosis into standard vocabulary. Agreement analysis.
Replicable finding due to (1) inclusion of diagnostic categories which are descriptive but not remediated; and (2) excessively complex diagnostic checklists	Observational Study emphasizing relationship between diagnostic and remediation plan. Agreement analysis based upon categories defined as problems and subject to remediation.

Replicable findings due to lack of (1) standard vocabulary (2) a clear diagnostic schema, and (3) a theoretic model coupled with practice in applying the model to diagnosis with feedback	Application Study to examine agreement improvements possible through brief (5 week) training program using materials developed in the IRT studies (e.g., the Model of Reading and Learning, Diagnostic Checklists, Simulated Cases, etc.)