Research Series No. 88

INVESTIGATIONS OF THE DIAGNOSTIC RELIABILITY OF READING SPECIALISTS, LEARNING DISABILITIES SPECIALISTS, AND CLASSROOM TEACHERS: RESULTS AND IMPLICATIONS

Annette B. Weinshank

Published By

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

September 1980

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 Program for Teaching and Instruction of the National Institute of Education, United States Department of Education. 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.

$2.00

Research Series No. 88

INVESTIGATIONS OF THE DIAGNOSTIC RELIABILITY OF READING SPECIALISTS, LEARNING DISABILITIES SPECIALISTS, AND CLASSROOM TEACHERS: RESULTS AND IMPLICATIONS

Annette B. Weinshank

Published By

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

Printed and Distributed
by
College of Education
Michigan State University

September 1980

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 Program for Teaching and Instruction of the National Institute of Education, United States Department of Education. 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

A study of reading specialists conducted in 1977 revealed that their diagnostic reliability was very low. Mean diagnostic agreement between two clinicians on statements seen as characterizing a case was effectively 0.00. Mean agreement for one clinician's diagnostic statements on a case and its replicate over time was less than 0.23. That is, fewer than one-quarter of the statements were repeated for the identical case. Five subsequent studies were designed to try and account for the low findings and to expand the generalizability of the results. Mean diagnostic agreement between two clinicians remained close to 0.00 across all five studies. Mean diagnostic agreement for one clinician on a case and its replicate remained close to 0.23 across the five cases. The results were consistent across fields (reading, learning disabilities), types of practitioner (clinicians, resource teachers, classroom teachers), and settings (laboratory and classroom).

INVESTIGATIONS OF THE DIAGNOSTIC RELIABILITY OF READING SPECIALISTS, LEARNING DISABILITIES SPECIALISTS, AND CLASSROOM TEACHERS: RESULTS AND IMPLICATIONS

Annette B. Weinshank[1]

The studies reported in this paper were designed by the Clinical Studies Group of the Institute for Research on Teaching to investigate the clinical problem-solving skills of reading and learning disabilities specialists and classroom teachers as they diagnosed, and then proposed remediations for, a variety of reading problems. The model for all these studies of clinician diagnostic performance was an Observational Study, conducted in 1977, of a selected group of eight highly trained and experienced reading specialists. The procedures developed for that study were used, with some modifications, (Gil, Hoffmeyer, Van Roekel, & Weinshank, Note 1) in all the subsequent investigations.

The 1977 Observational Study

Design and Analysis

Each of the eight clinicians in the 1977 study was randomly assigned to a simulated case of reading difficulty. A simulated case is a collection of information about a child with a reading problem. The four simulated cases developed by the Clinical Studies Group were based on real children in grades three through seven who had attended the Michigan State University Reading Clinic and were considered to be representative of reading problems commonly encountered in public schools.

---

[1]Annette B. Weinshank is a teacher co-investigator with the Clinical Studies Project and a former research intern with that project. She is an experienced reading specialist working with the Lansing, Michigan public schools. She holds a Ph.D. in Educational Psychology from Michigan State University.

Each simulated case was kept in a large file box. A cue inventory was provided listing the information (cues) available for each case: achievement tests, family and academic background, cognitive ability, group and individual reading diagnostic measures, classroom information, work samples, and so on. The information was presented in a variety of formats: test booklets, audio tapes, examiner's comments, and test scores. Each simulated case also had an equivalent form--a superficially disguised replicate of the original, prepared by making minor changes in the data base and randomly reordering the cue inventory (Lee & Weinshank, Note 2). In addition to these eight cases (four originals and four replicates), two learning disability cases were subsequently developed, and then two cases of reading difficulty that used only materials available in any ordinary classroom.

In all observational sessions, the task for each of the eight clinicians was to look at case information and write up a diagnosis and an initial remediation plan. Each clinician participated in three observational sessions, with sessions one and three separated by a minimum of one week. In the first session, the clinicians worked with one of the four original simulated cases. In the second session, they worked with one of the remaining three simulated cases. In the third session, they worked with the equivalent form (replicate) of the simulated case they had worked with in the first session. There were 24 sessions in all, six sessions per case.

During each observational session, the clinician requested items of information from the cue inventory. These were handed to the clinician one at a time by the experimenter, who recorded the cues requested. The experimenter managed the observational sessions by providing all necessary materials, timing the various tasks, and audio-taping the proceedings.

When the period allotted for cue collection was over, the clinicians were given a fixed amount of time to write out their diagnostic judgments and suggest an initial remediation plan. They were allowed to keep all the cues they had requested and any notes they had made while examining the cues.

The written diagnoses resulting from the observational sessions served as the unit of analysis for this study. The diagnoses were analyzed to determine (1) group agreement and (2) agreement between clinicians (inter-correlation) and between each clinician and him/herself (intracorrelation) on the diagnostic statements believed to characterize each case.

Extent of group agreement (the proportion of clinicians agreeing on statements seen as characterizing each case) was measured using a proportional agreement statistic. Since there were six sessions devoted to each case, any given statement could be mentioned from zero to six times. If a statement such as "instant sight word recognition low" was mentioned in three sessions, for example, the proportional agreement for that statement would be .50; that is, half the clinicians agreed that it characterized the case. If all clinicians across the six sessions agreed on a diagnostic statement, the proportional agreement would be 1.00.

To examine agreement between two clinicians on diagnostic statements seen as characterizing a case, or for one clinician on a case and its replicate, correlation matrices were constructed. A partial intra-clinician correlation matrix is presented in Table 1. A phi correlation was calculated for each matrix (Appendix A) as well as a second statistic, the Porter correlation, developed by Andrew Porter (Appendix A). All inter- and intracorrelations are presented using the phi correlation and, next to it in parenthesis, when available, the Porter statistic.

Table 1.

Partial Intracorrelation Matrix for Diagnosis
Case 3, Clinician A, Time One, Clinician
A, Time Two

STATEMENT ++

Present at Time One and Time Two

Motor coordination (W)
Intellectual potential:
General (S)
Oral reading: Phrasing (W)
Oral reading: Intonation (W)

STATEMENT +-

Present at Time One, Absent at Time Two

Attitude toward reading:
Independent (Obs)
Motivation for reading (Obs)
Emotional adjustment (W)
Substitutions contextually
acceptable (W)
Silent reading comprehension (W)
Word analysis (W)
Phonetic analysis (W)
Comprehension vocabulary (Obs)

STATEMENT -+

Absent at Time One, Present at Time Two

Motor coordination (Obs)
Hearing acuity (W)
Speech articulation (W, Obs)
Attitude toward reading:
Independent (W)
Attitude toward reading:
Instructional (W)
Relationship to peers (W)
Ability to apply reading
skills (W)
Oral reading: General (Obs)
Oral reading: Rate (W)
Oral reading:
Self-correction (W)
Silent reading: General (W)
Word analysis: General (S)
Phonetic analysis: General (Obs)
Use of initial consonant
sounds (S)
Use of syllables (S)
Word recognition: General (S)
Comprehension: Oral (S)
Comprehension: Listening (S)

STATEMENT --

Absent Both at Time One and Time Two

272 Domain statements excluded

Results

The results of the 1977 Observational Study (Vinsonhaler, Note 3) with respect to group agreement showed that most diagnostic statements were made only once for a given case. Across cases, only six diagnostic statements were mentioned in three or more sessions. These statements were: (1) at least average reading potential, (2) poor oral reading, (3) sight words low, (4) phonics weak, (5) poor word analysis skills, and (6) auditory discrimination problem. Despite lengthy individual diagnostic write-ups, the clinicians could agree on only a few statements characterizing any given case.

When the diagnostic statements of any two clinicians on a given case were compared, the results showed that on the average, they agreed on virtually no diagnostic statements. When the diagnostic statements across two cases (case/replicate) for a single clinician were compared, it was found that, on the average, fewer than one quarter of the statements mentioned by each clinician the first time s/he diagnosed a case were repeated when s/he diagnosed the replicate of the case.

The diagnostic agreement for the specialists in the 1977 observational study is summarized in Table 2.

Table 2.

Mean Diagnostic Agreement of Reading Specialists

	Intercorrelations		Intracorrelations	
	Phi	Porter	Phi	Porter
Observation Study, 1977	-0.07	(0.00)	0.13	(0.23)

The unexpectedly low diagnostic agreement results in this study were startling, particularly since the clinicians who participated were highly trained (all but two had doctoral degrees) and had an average of 10 years experience in their field. Clearly, before any conclusions or generalizations could be drawn, a number of possible explanations for the results had to be ruled out. The Clinical Studies group tested the validity of seven of these possible explanations, or hypotheses.

Hypothesis 1. The Findings Were Low Due to the Nature of the Training Reading Clinicians Received

Perhaps the clinicians were not adequately prepared to be consistent diagnosticians. Additionally, perhaps the sample was an unrepresentative one, and another group of similarly trained specialists would perform more reliably.

In order to test these hypotheses, a second observational study was conducted (VanRoekel, Note 4). Twenty learning disabilities clinicians and 20 reading clinicians diagnosed two simulated cases: One was a child with learning as well as reading disabilities; the other was a child with a reading disability only. If differences in training were indeed a key factor in performance, then the learning disabilities specialists could be expected to show greater agreement on the learning disability case, while the reading specialists should show greater agreement on the reading case.

The group agreement results paralleled those of the 1977 study. Despite lengthy individual diagnostic write-ups, a very small number of statements were agreed-upon as characterizing the learning disabilities case (weakness in gross/fine coordination; problem with visual perception/discrimination/memory/motor skills) and the reading case (average intellectual potential; problem with attitude/interests; weak phonic

analysis skills; observations about contextual reading ability).

The performance of the two groups of clinicians revealed no differential training effect whatever. Both groups performed at a near zero level of reliability, even within their own area of specialization (Table 3).

Table 3.

Mean Diagnostic Agreement of Reading and Learning Disabilities Specialists

	Intercorrelations			
	Reading Case		Learning Disability Case	
Observational Study of Reading and Learning Disabilities Specialists	Reading Specialists	Learning Disability Specialists	Reading Specialists	Learning Disability Specialists
	0.06	0.04	0.01	0.07
	$\bar{X} = 0.05$		$\bar{X} = 0.04$	

In sum, an average of only 5% of the diagnostic statements made for a case could be agreed-upon by any two clinicians examining that case.

These figures very nearly duplicated the intercorrelations reported for the 1977 study, and it was therefore felt that unique training effects and sampling error could be ruled out as explanations for the low diagnostic agreement of the reading specialists.

Hypothesis 2. The Findings Were Low Due to the Nature of Clinical Training in Reading and Learning Disabilities

Perhaps the training programs, both in reading and learning disabilities, were deficient in that they did not provide sufficient opportunity for the clinicians to implement their diagnostic findings in classroom

settings where they could get feedback about the accuracy of their diagnostic judgments. Classroom teachers, on the other hand, trained in using test-teaching in classroom settings might make more reliable diagnostic judgments.

This hypothesis was tested in a third observational study (Gil, Note 5). Ten classroom teachers participated in this study, five from the Lansing, Michigan area and five from the Chicago, Illinois area. Two classroom-oriented simulated cases were developed. The Chicago teachers had been trained to perform diagnoses using only materials normally available in a classroom; the Lansing teachers had not received such training.

Once again, only a small portion (6%) of total diagnostic statements made by the group were agreed upon as being characteristic of the cases: (1) poor comprehension, (2) knows major vocabulary concepts, (3) sight words weak, (4) ignores endings, (5) sight vocabulary good, (6) phonic skills weak, (7) problems with oral reading, and (8) word attack skills.

The extent of agreement between two teachers on diagnostic statements was effectively zero. There was no difference between teachers who had been trained in techniques of classroom diagnosis and those who had not (Table 4).

Table 4.

Mean Diagnostic Agreement of Classroom Teachers

	Intercorrelations	
	Case 7	Case 8
Observational Study of Classroom Teachers	- 0.04	- 0.03

Once again, the findings replicated those of earlier studies: (1) a meager consensus on statements characterizing a case was discerned only by aggregating diagnoses across clinicians; and (2) clinicians exhibited extremely low levels of agreement on the same case. Thus, ostensible differences in training programs did not result in differences in performance for the reading and learning disabilities specialists and classroom teachers investigated thus far. Perhaps these subjects did not reliably diagnose the cases because no program trained them to do so. However, other possible explanations for the low findings needed to be pursued.

Hypothesis 3. The Findings Were Low Due to a Lack of Standardized Vocabulary Among the Clinicians

Perhaps in categorizing the clinicians' natural language statements, the experimenters failed to see equivalences. In that case, statements that were actually describing the same thing would be coded as being dissimilar, and agreement would appear to be very low.

A fourth observational study was undertaken (Hoffmeyer, 1980). This study was designed to replicate the initial 1977 investigation, adding the use of a standardized diagnostic checklist empirically derived from clinicians' statements in the preceding observational studies. The reading clinicians transferred their own natural language statements to the standardized checklist. All analyses were based on the checklists, thereby eliminating all coder subjectivity with respect to equating natural language statements.

Group agreement across cases focused on the same diagnostic categories as the original investigations: (1) at least average intellectual potential, (2) poor oral reading, (3) sight words low, (4) phonics weak, (5) poor word analysis skills, and (6) auditory acuity. There was some

agreement on two additional categories: problem with comprehension and poor attitude toward reading. For any given case, then, only a few statements could be gleaned which represented group agreement on case characteristics.

The results of comparing the diagnostic statements of any two clinicians for a case showed a slight increase over the original study. The performance of one clinician over time remained essentially the same (Table 5).

Table 5.

Mean Diagnostic Agreement of Reading Specialists

	Diagnostic Agreement			
	Intercorrelations		Intracorrelations	
	Phi	Porter	Phi	Porter
1978 Replicate using Standardized Diagnostic Checklist	0.11	(0.11)	0.25	(0.17)

Thus, differences in vocabulary did not appear to change clinician reliability to any significant degree.

Hypothesis 4.
The Findings Were Low Due to the Nature of the Experimental Setting

Perhaps the requirement that information be requested item by item was too dissimilar from the conditions under which clinicians actually went about the task of diagnosing reading difficulties.

A fifth observational study (Stratoudakis, Note 7) was carried out in which the simulated case format was altered. The amount of case material was reduced, and the information was presented in a three-ring notebook. Instead of requesting items of information from an experimenter,

the clinicians worked independently with all the cues at hand in the notebook. (This format had the additional advantage of being more economical: fewer items of information reduced materials costs and an experimenter was no longer required.)

Diagnostic agreement both between and within clinicians replicated that found in earlier studies (Table 6). Thus, diagnostic agreement remained virtually unchanged despite the altered format and procedures.

Table 6.

Mean Diagnostic Agreement Using Notebook Format

	Diagnostic Agreement			
	Intercorrelations		Intracorrelations	
	Phi	Porter	Phi	Porter
Observational Study Using Notebook Format	0.13	(0.10)	0.20	(0.12)

A final set of hypotheses was formulated in an attempt to account for the low diagnostic findings that had continued to be replicated through the preceding studies. These hypotheses were tested in a final observational study that emphasized the relationship between diagnosis and remediation.

Hypothesis 5.
The Findings Were Low Due to the Inclusion of Descriptive Diagnostic Statements

Perhaps we in the Clinical Studies group were overlooking a core group of diagnostic statements to which remediations were consistently attached. By using as the unit of analysis all diagnostic statements made instead of just those which were seen as needing remediation, we might have inadvertently "swamped" substantial agreement on remediated diagnostic statements.

Hypothesis 6. The Findings Were Low Because the Checklist Used in the 1978 Replication Study Might Have Been Excessively Complex

A shorter more tightly organized checklist derived from that earlier study might make the translation process more accurate.

Hypothesis 7. The Findings Were Low Because Diagnostic Reliability Is Not Important to Clinicians But Remedial Reliability Is and Will Be Reflected in Greater Reliability with Respect to Actions Chosen

The final observational study (Weinshank, Note 8) in this three-year series of investigations addressed Hypotheses 5, 6, and 7. In this study, the eight experienced reading clinicians, four trained in Michigan and four in Illinois, transferred their diagnostic statements to a shortened diagnostic checklist, their remedial statements to an empirically derived remedial checklist, and explicitly associated remedial and diagnostic statements.

Yet again, a small portion (10%) of diagnostic categories mentioned accounted for whatever group agreement existed across cases: (1) at least average intellectual potential, and problems with (2) word recognition, (3) word analysis, (4) oral reading, (5) silent reading, (6) comprehension, (7) auditory/visual acuity, (8) auditory discrimination, and (9) affect. The results are summarized in Table 7.

Table 7.

Mean Agreement for Diagnosis, Remediation, and Remediated Diagnoses

		Agreement			
		Intercorrelations		Intracorrelations	
		Phi	Porter	Phi	Porter
Observational Study Emphasizing Relationship Between Diagnosis and Remediation	Diagnosis	0.16	(0.11)	0.23	(0.14)
	Remediation	0.14	(0.10)	0.29	(0.20)
	Remediated Diagnoses	0.13	(0.08)	0.22	(0.14)

The results showed that (1) global diagnostic reliability remained unacceptably low; (2) agreement on remedial actions to be used fared equally poorly; and (3) agreement on precisely which diagnoses warranted treatment fell lower still.

Thus, none of the final three hypotheses was supported by the data from this study.

Conclusions

Low diagnostic (and remedial) reliability for reading and learning disabilities specialists and classroom teachers appears to be a robust phenomenon (Table 8). Across studies, the mean agreement between any two clinicians on a given case was 0.08 (the range was 0.00 to 0.16). They agreed, on the average, on only 8% of their combined statements, which is no better than the agreement expected due to chance. The mean agreement across studies for a single clinician on a case and its replicate was 0.20 (the range was 0.13 to 0.25). Given the identical case on two separate occasions, a given clinician, on the average, agreed with him/herself on only 20% of the combined statements for the case and its replicate.

Why are these trained and experienced professionals performing so unreliably? Can their diagnostic and remedial reliability be improved?

Further analyses of individual performance in the studies reported here seemed to confirm that the chief cause for low diagnostic and remedial reliability in reading was inadequate or inappropriate training (Vinsonhaler, Note 3).

Short-term (30 instructional hours) training studies conducted by the Clinical Studies group (Sherman, Weinshank, & Brown, Note 9; Gil, Polin, Vinsonhaler, & VanRoekel, Note 10) have pilot-tested procedures which doubled the average entering reliability of those who were trained.

Table 8.

Diagnostic Reliability of Reading Specialists, Learning Disabilities Specialists and Classroom Teachers

	Intercorrelations		Intracorrelations	
	Phi	Porter	Phi	Porter
Observational Study, 1977	-0.07	(0.00)	0.13	(0.23)
Observational Study of Reading and Learning Disability Specialists	0.05 (reading case)	--	--	--
	0.04 (Learning Disability Case)	--	--	--
Observational Study of Classroom Teachers	-0.04 (Case 7)	--	--	--
	-0.03 (Case 8)	--	--	--
Observational Study 1977 Replication	0.11	(0.11)	0.25	(0.17)
Observational Study Using Notebook Format	0.13	(0.10)	0.20	(0.12)
Observational Study Emphasizing Relationship Between Diagnosis and Remediation Plan	0.16 (Diagnosis)	(0.11)	0.23	(0.14)
	0.14 (Remediation)	(0.10)	0.29	(0.20)
	0.13 (Remediated diagnoses)	(0.08)	0.22	(0.14)

The training included: (1) the use of a model-based procedure to guide the tasks of diagnosis and prescription, (2) direct training in the use of decision aids and standard vocabulary, and (3) extended practice with feedback on diagnostic performance.

These studies have not, however, ruled out other possible sources of improvement, nor have they determined how, or whether, more substantial gains might be achieved. Research now in progress in the Clinical Studies Project is addressing these questions. Additionally, the validity of model-directed diagnosis and remediation is being studied in the context of systematic classroom-based follow-up of children with reading difficulties. Taken together, these lines of inquiry are likely to have significant implications for preservice teacher training, advanced clinical training, and inservice professional development programs.

Reference Notes

1. Gil, D., Hoffmeyer, L., Van Roekel, J., Vinsonhaler, J., & Weinshank, A. Clinical problem solving in reading: Theory and research (Res. Ser. No. 45). East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1979.

2. Lee, A., & Weinshank, A. CLIPIR pilot observational study of reading diagnosticians (Res. Ser. No. 14). East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1978.

3. Vinsonhaler, J. The consistency of reading diagnosis (Res. Ser. No. 28). East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1979.

4. Van Roekel, J. The problem-solving behavior of reading specialists and learning disabilities resource teachers. Doctoral dissertation in progress, Michigan State University.

5. Gil, D. The clinical problem-solving behavior of classroom teachers as they diagnose a child's reading behavior in laboratory and classroom. Unpublished doctoral dissertation, Michigan State University, 1979.

6. Hoffmeyer, L. The process and outcomes of diagnostic problem solving among reading clinicians. Unpublished doctoral dissertation, Michigan State University, 1980.

7. Stratoudakis, J. The construction and use of paper problems to observe the diagnostic problem solving behavior of reading clinicians. Unpublished doctoral dissertation, Michigan State University, 1980.

8. Weinshank, A. An observational study of the relationship between diagnosis and remediation in reading. Doctoral dissertation, Michigan State University, 1980. (Summarized in IRT Res. Ser. No. 72, East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1980.)

9. Sherman, G., Weinshank A., & Brown, S. Training reading specialists in diagnosis (Res. Ser. No. 31). East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1978.

10. Gil, D., Polin, R., Vinsonhaler, J., & Van Roekel, J. The impact of training on diagnostic consistency (Res. Ser. No. 67). East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1980.

APPENDIX A

Calculation of Phi Correlation and Porter Statistic

Clinician i SIMCASE Q, Form One

Clinician i SIMCASE Q, Form Two	PRESENT (+)	ABSENT (-)
	Frequency count of statements present in the domain in both sessions for form one and form two of SIM-CASE **A**	Frequency count of statements present in the domain present in SIMCASE form two but not in SIMCASE form one **B**
	Frequency count of statements in the domain present in the session for SIM-CASE form one but not SIMCASE form two **C**	Absent in both sessions for form one and form two of SIMCASE **D**

	+	-	
+	a (++)	b (+-)	a + b
-	c (-+)	d (--)	c + d
	a + c	b + d	N

$$\text{Phi} = \frac{(a \times d - b \times c)}{(a + C) \times (b + d) \times (c + d) \times (a + b)}$$

The presence of a large percentage of statements (more than 85%) in the "D" cell (the statement is absent in both sessions) artificially inflated the intercorrelations since it represented, in effect, agreeing to disagree. A statistic developed by Professor A. Porter (Institute for Research on Teaching, Michigan State University) was designed to correct for this occurrence, by including in the computation only the values in the A, B, and C cells $\left(\frac{A}{A + B + C}\right)$.