

# CANADIAN THESES ON MICROFICHE

## THÈSES CANADIENNES SUR MICROFICHE



National Library of Canada  
Collections Development Branch

Canadian Theses on  
Microfiche Service

Ottawa, Canada  
K1A 0N4

Bibliothèque nationale du Canada  
Direction du développement des collections

Service des thèses canadiennes  
sur microfiche

### NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

**THIS DISSERTATION  
HAS BEEN MICROFILMED  
EXACTLY AS RECEIVED**

### AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

**LA THÈSE A ÉTÉ  
MICROFILMÉE TELLE QUE  
NOUS L'AVONS REÇUE**

**Canada**



National Library of Canada / Bibliothèque nationale du Canada

Ottawa, Canada / K1A 0N4

TC -

ISBN 0-315-21199-7

CANADIAN THESES ON MICROFICHE SERVICE - SERVICE DES THÈSES CANADIENNES SUR MICROFICHE

PERMISSION TO MICROFILM - AUTORISATION DE MICROFILMER

• Please print or type - Écrire en lettres moulées ou dactylographier

AUTHOR - AUTEUR

Full Name of Author / Nom complet de l'auteur

Meilicke, Dorothy T.

Date of Birth - Date de naissance

April 6 / 1936

Canadian Citizen - Citoyen canadien

Yes / Oui

No / Non

Country of Birth - Lieu de naissance

Canada

Permanent Address - Résidence fixe

13907-78 Ave, Edmonton  
Alberta

THESIS - THÈSE

Title of Thesis - Titre de la thèse

Variables Associated with the Integrated  
Reliability of a Quality Monitoring Instrument.

Degree for which thesis was presented /  
Grade pour lequel cette thèse fut présentée

MED

Year this degree conferred /  
Année d'obtention de ce grade

75

University - Université

Queen's

Name of Supervisor - Nom du directeur de thèse

AUTHORIZATION - AUTORISATION

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to  
microfilm this thesis and to lend or sell copies of the film.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE  
DU CANADA de microfilmer cette thèse et de prêter ou de vendre des ex-  
emplaires du film.

The author reserves other publication rights, and neither the thesis nor exten-  
sive extracts from it may be printed or otherwise reproduced without the  
author's written permission.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs ex-  
traits de celle-ci ne doivent être imprimés ou autrement reproduits sans  
l'autorisation écrite de l'auteur.

ATTACH FORM TO THESIS - VEUILLEZ JOINDRE CE FORMULAIRE À LA THÈSE

Signature

*D. Meilicke*

Date

April 11/85

THE UNIVERSITY OF ALBERTA

VARIABLES ASSOCIATED WITH THE INTERRATER RELIABILITY  
OF A QUALITY MONITORING INSTRUMENT

by

DOROTHY MEILICKE

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE  
OF MASTER OF EDUCATION

DEPARTMENT OF EDUCATIONAL ADMINISTRATION



EDMONTON, ALBERTA

SPRING, 1985



14 February 1985

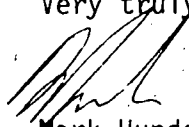
Ms. Dorothy Meilicke  
University of Alberta Hospitals  
112 Street & 84 Avenue  
Edmonton, Alberta  
T6G 2B7

Dear Dorothy,

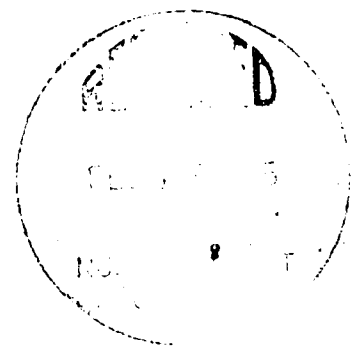
This letter will formally grant you permission to include elements of the Medicus/RPSL Nursing Quality Monitoring System in your thesis. We understand that you will be including copies of questionnaires and answer sheets in the appendix to your thesis.

I am looking forward to the opportunity to review your work. I hope that we have an opportunity to see each other sometime soon.

Very truly yours,

  
Mark Hundert  
President

MH/cm



THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR

DOROTHY MEILICKE

TITLE OF THESIS

VARIABLES ASSOCIATED WITH  
THE INTERRATER RELIABILITY  
OF A QUALITY MONITORING  
INSTRUMENT

DEGREE FOR WHICH THESIS WAS PRESENTED

M.Ed.

YEAR THIS DEGREE GRANTED

1985

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive abstracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed)

*D. Meilicke*

PERMANENT ADDRESS:

*13907-78 Ave,*

*Edmonton, Alberta*

DATED

*April 11*

1985

THE UNIVERSITY OF ALBERTA  
FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Variables Associated With the Interrater Reliability of a Quality Monitoring Instrument submitted by Dorothy Mellicke in partial fulfilment of the requirements for the degree of Master of Education

*Dona Fay*  
.....  
Supervisor  
*Heide Coker*  
.....  
*A. Horn*  
.....

Date *April 3/85* .....

## ABSTRACT

This study examined the relationship of observer specialization and of data collection methods to interrater reliability when using a quality monitoring instrument. Observer specialization was defined as the degree nurse observers conduct observations in their area of current clinical experience. Data collection methods were defined as the different sources of information specified in the Rush-Medicus Quality Monitoring Instrument; namely, interview, observation, inference, and patient record abstraction. Several null hypotheses were tested:

- 1 There are no differences in the interrater reliability scores between specialized and nonspecialized observers.
- 2 There are no differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference and record abstraction methods.
- 3 There are no differences in interrater reliability scores for data collected by each method within either the specialized or nonspecialized observer groups.

The methodology employed the Posttest - Only Control Group design. A total of 62 subjects consented to participate in the study. Thirty-one pairs of randomly selected nurse observers were assigned to either control or experimental group status. The control pairs each conducted one interrater reliability testing observation on a randomly selected nursing unit that matched the nurse observer's current clinical experience. The experimental pairs each conducted one

interrater reliability testing observation on a randomly selected nursing unit that did not match the observer's current clinical experience. Following each pair's interrater reliability observation, the reasons for disagreement were discussed and the percentage agreement for each pair was calculated.

The results derived from the statistical analyses of the data suggest, in regard to observer specialization, that the nonspecialized observers achieved significantly higher interrater reliability estimates than did the specialized observers. This unexpected finding is subject to a number of possible explanations, not the least of which is the possibility that an "reactive arrangements" effect occurred within the experimental group.

In regard to data collection methods, the findings suggest that the patient record abstraction method presents the most serious difficulty to nurse observer's using the Rush-Medicus Quality Monitoring Instrument. Furthermore, this finding implies that data obtained from the patient record ought to be viewed with caution until acceptable levels of reliability are achieved.

In respect to reasons for disagreement, the finding that 58 - 66 percent of the total disagreements were related to judgement varies and observer carelessness suggests that factors such as observer selection, tenure, retraining, and frequency of interrater reliability testing need to be seriously addressed when using an ongoing data collecting instrument. Further research on these factors is needed.



## ACKNOWLEDGEMENTS

The author wishes to express sincere gratitude to the following individuals, who have contributed to the completion of this project:

Dr. D.A. MacKay, for his interest, direction, and encouragement as advisor on the thesis;

Dr. A. Konrad and Dr. R. Cockerill for their helpful comments, and participation on the examining committee;

Mrs. C. Prokop, for her creativity and willing assistance in the analyses of the data;

Mrs. D. Gitzel and Ms. L. Laird for their commitment and time in typing the manuscript;

Nurse observers at the study hospital for their valuable participation in the study;

My friends and colleagues in Nursing Administration, University of Alberta Hospitals, for their continued encouragement and interest throughout the project; and,

My family -- Carl, Michelle, Michael, and Jackie -- for their years of patience, understanding, and support in the completion of this project.

## TABLE OF CONTENTS

CHAPTER	PAGE
I INTRODUCTION . . . . .	1
Background to the Problem . . . . .	1
Problem Statement . . . . .	2
Definition of Terms . . . . .	3
Significance of the Study . . . . .	4
Survey of hospitals . . . . .	5
Observer specialization . . . . .	8
Data collection methods . . . . .	9
Description of the Variables . . . . .	10
Independent variables . . . . .	10
Dependent variable . . . . .	11
Hypotheses . . . . .	11
Hypothesis 1 . . . . .	11
Hypothesis 2 . . . . .	11
Hypothesis 3 . . . . .	12
Limitations and Delimitations . . . . .	12
Limitations . . . . .	12
Delimitations . . . . .	14
Organization of the thesis . . . . .	14
II LITERATURE REVIEW . . . . .	16
Reliability of Measurement Instruments . . . . .	16
Factors Affecting Measurement Error . . . . .	16
Methods of Assessing Reliability . . . . .	22
Assessment of Interrater Reliability . . . . .	23
III METHODOLOGY . . . . .	26
Type of Study . . . . .	26
Setting . . . . .	26
Population of Specialized Observers . . . . .	28
Sampling Plan . . . . .	29
Random Sampling of Subjects . . . . .	29
Random Assignment of Subjects to Pairs . . . . .	30
Random Assignment of Pairs to Experimental or Control Group . . . . .	31
Random Assignment of Pairs to Nursing Units . . . . .	31
Instrumentation . . . . .	32
Data Collection Methods . . . . .	33
Measurement of the Rush-Medicus Instrument . . . . .	33
Reliability of the Rush-Medicus Instrument . . . . .	34
Validity of the Rush-Medicus Instrument . . . . .	36
Study Design . . . . .	37

	Ethical Considerations . . . . .	38
	Data Collection Procedure . . . . .	39
	Data Analyses . . . . .	40
IV	RESULTS . . . . .	48
	Interrater Reliabilities Estimated by	
	Percentage Agreement . . . . .	48
	Percentage agreement by observation . . . . .	48
	Percentage agreement by objective . . . . .	49
	Percentage agreement by source of information . . . . .	51
	Significant Differences and Hypotheses Testing . . . . .	54
	All observations combined . . . . .	54
	By objective . . . . .	54
	By source of information . . . . .	56
	Sources within each group . . . . .	56
	Reasons for Disagreements . . . . .	61
	Reasons as a percent of total disagreement . . . . .	61
	Differences between groups' reasons . . . . .	63
	Differences within group's reasons . . . . .	63
	Objective 2 (Physical Needs Are Attended) . . . . .	67
	Objective 2 by subobjective . . . . .	67
	Objective 2 by source of information . . . . .	70
	Reasons within Objective 2 . . . . .	73
V	SUMMARY, DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS . . . . .	74
	SUMMARY . . . . .	74
	Percentage agreement . . . . .	80
	Hypotheses testing . . . . .	81
	Reasons for disagreements . . . . .	84
	DISCUSSION . . . . .	88
	Similar Findings Between the Groups . . . . .	88
	All objectives combined . . . . .	88
	By objective . . . . .	89
	By source . . . . .	90
	By reason . . . . .	91
	Significant Differences Between the Groups . . . . .	93
	Ho 1 . . . . .	93
	Ho 2 . . . . .	94
	Ho 3 . . . . .	95
	Significant differences among reasons . . . . .	96
	Limitations . . . . .	100

**CHAPTER**

**PAGE**

Implications . . . . . 101  
    Specialized versus nonspecialized observers . . . . . 101  
    Interrater reliability by objective . . . . . 101  
    Interrater reliability by source . . . . . 102  
    Reasons for disagreement . . . . . 102  
CONCLUSIONS AND RECOMMENDATIONS . . . . . 105  
    Observer Specialization and Interrater Reliability . . . 105  
    Recommendations for further research . . . . . 105  
    Data Collection Methods and Interrater Reliability . . . 106  
    Recommendations for further research . . . . . 106  
    Reasons for Disagreements . . . . . 106  
    Recommendations for further research . . . . . 107  
REFERENCES . . . . . 109  
APPENDIX . . . . . 111

## LIST OF TABLES

TABLE		PAGE
1	Interrater Reliability for each Observation Estimated by Percent Agreement . . . . .	49
2	Interrater Reliability by Objectives Estimated by Percent Agreement . . . . .	50
3	Interrater Reliability by Source Estimated by Percent Agreement . . . . .	52
4	Differences Between Groups by Objectives Estimated by Chi-Square Analyses . . . . .	55
5	Differences Between Groups by Source Estimated by Chi-Square Analyses . . . . .	57
6	Differences Within Experimental Group's Sources Estimated by Chi-Square Analyses . . . . .	59
7	Differences Within Control Group's Sources Estimated by Chi-Square Analyses . . . . .	60
8	Reasons for Disagreements Estimated by Percentage . . . . .	62
9	Differences Between Groups' Reasons Estimated by Chi-Square Analyses . . . . .	64
10	Differences within Experimental Group's Reasons Estimated by Chi-Square Analyses . . . . .	66
11	Differences within Control Group's Reasons Estimated by Chi-Square Analyses . . . . .	68
12	Differences Between Groups by Subobjective (Objective 2) Estimated by Chi-Square Analyses . . . . .	69
13	Differences Between Groups' by Sources (Objective 2) Estimated by Chi-Square Analyses . . . . .	71
14	Differences Between Groups' by Reasons (Objective 2) Estimated by Chi-Square Analyses . . . . .	72

## CHAPTER I

### INTRODUCTION

In Canadian health care organizations, there is increasing use of data collecting instruments which require the use of raters, observers, or interviewers. This is a result of a growing concern about, and attention to, quality assurance activities. In Canadian hospitals, for example, this concern comes not only from the recent guidelines of the Canadian Council on Hospital Accreditation, which stipulate that each hospital must monitor the quality of services provided, but also from an increasing pressure to manage resources in ways which are both morally and legally defensible (CCHA, 1977, 1983; Mickevisius and Stoughton, 1984). Given the foregoing, it is of utmost importance to achieve acceptable levels of reliability with these instruments, but the fact is that minimal research has been done on the reliability, or on the factors affecting the reliability, of these quality monitoring instruments (Hegyvary and Hausmann, 1976; Ventura, Hageman, Slakter and Fox, 1982).

#### Background to the Problem

Reliability is a basic attribute that every data collecting instrument (measure) must possess if it is to be useful (Fox, 1982). Reliability is the degree to which a measure yields consistent, accurate responses (Kerlinger, 1973). The purpose of establishing reliable measures is to reduce response error: that portion of an actual response that varies from the true response (Kerlinger, 1973; Sudman and Bradburn, 1974).

There are a multitude of factors that might affect the size and

direction of response error. For data collecting instruments which utilize raters, factors relating to the instrument, the rater (observer, interviewer), and the respondent should be considered in order to reduce response error (Kahn and Cannell, 1963; Sudman and Bradburn, 1974; Kerlinger, 1973; Giovannetti, 1981). Accordingly, for data collection instruments which require the use of raters, observers, or interviewers, the reliability of the instrument becomes a function of the instrument itself, plus the rater, observer or interviewer, and the respondent (Polit and Hungler, 1978; Sudman and Bradburn, 1974). For instruments of this nature, the most critical reliability assessment method is interrater reliability (Ventura et al. 1980). Interrater reliability is an estimate of the degree to which two or more independent raters, observers, or interviewers are consistent in their judgements (Goodwin and Prescott, 1981, p. 324). This is a particularly difficult problem because the reliability of an instrument is not a fixed property, but instead it is a condition of the instrument which needs to be established in each setting where the instrument is being used (Polit and Hungler, 1978; Ventura et al. 1980).

In this study, the relationship of observer specialization and of data collection methods to interrater reliability when using a quality monitoring instrument was examined.

### Problem Statement

The purpose of this study was to investigate the following three questions regarding interrater reliability of a quality monitoring instrument:

1. What is the relationship that exists between observer specialization and interrater reliability when using a quality monitoring instrument?
2. What is the relationship that exists between different data collection methods and interrater reliability when using a quality monitoring instrument?
3. What is the relationship that exists between different data collection methods and interrater reliability when observer specialization is controlled for?

#### Definition of Terms

The specific meaning of key terms and concepts used in this study are listed below:

**Nurse observers.** Nurses who are trained to conduct quality monitoring observations using the Rush-Medicus Quality Monitoring Instrument.

**Specialized observers.** Nurse observers who conduct observations only in their area of current clinical experience; for example, surgical nurse observers observing in surgical nursing areas.

**Nonspecialized observers.** Nurse observers who conduct observations in areas in which they do not have current clinical experience; for example, surgical nurse observers observing in pediatric areas.

**Control group.** Pairs of specialized observers who conduct observations in their areas of specialization.

**Experimental group.** Pairs of nonspecialized observers who conduct observations in areas in which they are not specialized.

**Observation.** An assessment of two randomly selected patients and



one nursing unit using the Rush-Medicus Quality Monitoring Instrument. This is commonly referred to as a nursing audit.

**Interrater reliability.** The reliability of an instrument as estimated by either the equivalence or percentage agreement approach. Interrater reliability is estimated by having two or more observers conduct the audit simultaneously and independently record their responses (Polit and Hungler, 1978).

**Significance of the Study**

In nursing departments of Canadian acute care hospitals, increasing attention is being given to the provision of objective measures of quality of nursing care. This attention results from a variety of pressures, among the most important of which are concerns about the effect on nursing care of staff reductions due to budgetary constraints and concerns over legal liability in the face of an increased incidence of malpractice litigation.

The development and use of such objective measures is of relatively recent origin and, consequently, many questions regarding the utility of the existing measures have yet to be answered. One of these questions has to do with reliability because, although there are instruments which are considered to be reasonably valid, few studies of their reliability or of the factors affecting their reliability have been done (Hegyvary and Hausmann, 1976; Ventura et al. 1980).

As mentioned previously, there are many factors that may affect the reliability of data collecting instruments which require the use of raters. Because of the ongoing nature of these quality monitoring

processes (audits) in nursing departments, a number of these factors become particularly relevant. These include:

1. the numbers of trained observers;
2. the observer's position in the organization;
3. the term of the observer's appointment;
4. the retraining needs of observers;
5. the frequency of interrater reliability testing;
6. the assignment plan of observers; and,
7. the relevance of question content over time.

Despite the lack of research on these factors, nursing administrators make, and are continuing to make, operational decisions regarding these factors each time a quality monitoring instrument is implemented. These decisions have led to considerable variations in practice. Examples of the differences in practice were noted by the researcher in a survey of hospitals conducted in conjunction with this study.

Survey of hospitals. In October, 1983, a questionnaire survey was sent to all Canadian acute care hospitals that use a similar nursing audit process -- the Rush-Medicus Quality Monitoring Instrument (Rush-Medicus) implemented by the Rush-Medicus Consulting Firm. The questionnaire was sent to 20 nursing departments across Canada. The demographic characteristics of these hospitals confirm their acute care status, (Appendix A). A 100 percent response rate was obtained. For these hospitals, the length of time since implementation of the Rush-Medicus Quality Monitoring Instrument ranged from less than 12 months (7 hospitals) to over 4 years (4 hospitals), with a mode of less than 12 months.

The survey questions focused primarily on factors associated with observers such as:

1. numbers trained;
2. position in the organization;
3. assignment plan;
4. training and retraining; and,
5. interrater reliability testing.

In addition, one question dealt with the structure in place to maintain the nursing audit (see Appendix B for the complete questionnaire).

In regard to the numbers of trained observers in different hospitals, the results of the survey indicated that the numbers vary widely among hospitals. The range was from less than 10 observers (one hospital) to over 91 observers (two hospitals), with a mode of 21 - 30 observers. Similarly, on the subject of the observer's position in the organization, the results indicated that nurses of many differing positions function as observers. Eighty-five percent of the hospitals reported charge nurses as observers, 45 percent reported supervisors, 40 percent reported using general duty nurses, 35 percent reported clinical instructors, and 20 percent reported assistant charge nurses. As well, 60 percent of the survey hospitals reported an "other" category for nurse observers.

On the topic of assignment plan for observers, a variation in how nurse observers are assigned was reported. Eighty percent of the hospitals indicated their nurse observers were assigned to audit on any nursing unit, whereas 20 percent of the hospitals indicated

their nurse observers were assigned only to nursing units that matched the observer's current clinical experience.

In regard to the training process for nurse observers, a variation in both the process and length of training was reported. Fifty percent of the hospitals reported a training process which consisted of a classroom review of the standard orientation manual, plus a practice session on a nursing unit, followed by an interrater reliability test. Of the remaining hospitals, 35 percent excluded the practice session, 05 percent excluded the interrater reliability test, 05 percent reported a three day workshop, and 05 percent, due to recency of implementation, had not yet conducted training sessions. On the topic of the length of training process, a variation was also reported. Forty percent of the hospitals reported a training process of 5 - 8 hours in length, 30 percent reported a process requiring 9 - 16 hours, 15 percent reported a process requiring less than 3 hours, and 10 percent reported a process requiring over 16 hours.

In regard to the subject of retraining sessions, 80 percent of the hospitals reported that retraining sessions were not conducted. Of the 20 percent that did conduct retraining sessions, the sessions were designed primarily to retrain observers after periods of absence from auditing.

In relation to the frequency of interobserver reliability testing after the initial orientation period, the findings again showed variations. Fifty-five percent of the hospitals reported conducting interrater reliability testing on a regular basis and 45 percent did not. Of the eleven hospitals conducting regular interrater reliability

testing, 10 percent conducted quarterly tests, 25 percent conducted semiannual tests, and 20 percent conducted annual tests. Ten of the eleven hospitals conducting interrater reliability testing reported the percentage agreement scores obtained in their reliability testing. Their mean score was 90 percent, mode 95 percent, and median 88 percent.

Finally, on the topic of a committee to support the ongoing maintenance and development of the quality monitoring process, 75 percent of the hospitals reported a special committee established for this purpose.

The foregoing description of the differences reported by twenty hospitals, all using a similar instrument, represents factors that may affect the reliability of an ongoing quality monitoring instrument. The above differences in practice need to be studied. In general, this study has been undertaken as a beginning examination of these factors. In particular, this study was undertaken to determine the relationship which exists between (a) observer specialization and interrater reliability, and (b) data collection methods and interrater reliability when using a quality monitoring instrument designed to measure quality of nursing care.

**Observer specialization.** In regard specifically to observer specialization and interrater reliability, the significance of the study is that many institutions now use nurses to observe in many different clinical areas whilst other hospitals only use nurses to observe in clinical areas in which the observer has current clinical experience. The effect on interrater reliability of this variation in observer specialization has not been studied. It could be argued

that the reliability of the observation will be lower if the nurse observers are not specialized in the area they are observing. One is attracted to this conclusion because it seems logical that, given the high degree of technical complexity in contemporary nursing practice, a nonspecialized observer will have a higher probability of inconsistent measurement in an unfamiliar clinical area than will a specialized observer (such as a pediatric nurse observing in intensive care units as compared to pediatric units). Furthermore, this argument tends to be supported by the few studies done on the reliability of quality monitoring instruments in which the researchers specifically selected nurses with experience in the clinical service area in which the study was conducted (Ventura et al. 1980; Hegyvary and Hausmann, 1976).

Data collection methods. In regard to data collection methods and interrater reliability, the significance of this study is that many quality monitoring instruments require more than one method of data collection. For example, the Quality Patient Care Scale (QualPacs) requires that data be collected by both direct observation and patient record abstraction (Ventura et al. 1980); in the same vein, the Rush-Medicus Quality Monitoring Instrument (Rush-Medicus) requires that data be collected by interview, observation, patient record abstraction, and observer inference. It is recognized in the social sciences literature that each of these methods has varying influences upon the reliability of an instrument but, unfortunately, the topic has received minimal attention in health services research with respect to the reliability of quality monitoring instruments (Herman and

Cayten, 1980; Kidder, 1981). One notable exception is the study by Herman and Cayten (1980) which found low interrater reliability on medical record abstraction was associated with variables which require the rater to use judgement. Recent personal observation by the researcher tends to support the above finding. The researcher has frequently noted that interrater agreement, for nurse observer trainees, appears to be lower for those variables which require the observer trainee to collect data via the patient's record. In addition, it appears that data collected by the interview method tends to have higher interrater agreement for nurse observer trainees.

Given the importance to both the hospital and to the patient of objective data regarding quality of nursing care, it is obviously important that the highest possible levels of reliability be achieved.

#### Description of the Variables

This research focused on the relationship between independent and dependent variables which are defined as follows:

##### Independent variables.

1. Observer specialization - defined as the degree to which nurse observers conduct observations (audits) in their areas of current clinical experience. This variable consisted of two elements: specialized observers and nonspecialized observers.
2. Data collection methods - defined as the source of information specified for each criterion in the Rush-Medicus Quality Monitoring Instrument. This variable consisted of eight elements: patient record abstraction, patient interview, nursing personnel interview, patient observation, nursing personnel observation, environmental

observation, unit management observation, and observer inference. These eight elements (sources of information) comprise four different data collection methods: interview, observation, inference, and record abstraction.

### Dependent variable.

1. Interrater reliability - defined as the degree to which two or more independent raters, scorers, judges, or interviewers are consistent in their judgements (Goodwin and Prescott, 1981, p. 325). This variable consisted of one measure: percentage agreement.

### Hypotheses

In this study the following hypotheses, described in both the null (Ho) and alternate (Ha) forms were tested:

#### Hypothesis 1.

Ho: There are no differences in the interrater reliability scores between specialized and nonspecialized observers.

Ha: There are differences in the interrater reliability scores between specialized and nonspecialized observers.

#### Hypothesis 2.

Ho: There are no differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference, and record abstraction methods.

Ha: There are differences between the interrater reliability scores of specialized and nonspecialized observers for data collected



by the interview, observation, inference, and record abstraction methods.

### **Hypothesis 3.**

Ho: There are no differences in interrater reliability scores for data collected by the interview, observation, inference, or record abstraction methods within either the specialized or the nonspecialized observer groups.

Ha: There are differences in interrater reliability scores for data collected by the interview, observation, inference, or record abstraction methods within either the specialized or the nonspecialized observer groups.

### **Limitations and Delimitations**

**Limitations.** In this study, the following limitations were identified:

1. Generalizability is limited because the random selection of nurse observers at the study hospital may not represent the general population of nurse observers at other hospitals.
2. Random assignment of nurse observers to the Posttest - Only Control Group design was expected to provide an adequate assurance of lack of bias between groups (Campbell and Stanley, 1963). However, there exists a slight chance that randomization failed and the groups would differ even if no treatment intervened (Kidder, 1981).
3. The Posttest - Only Control Group design controls for factors jeopardizing internal validity such as history, maturation, testing, instrumentation, regression, and selection bias. However, experimental mortality (i.e., differential loss of subjects)

- must be considered in the data analyses procedures (Campbell and Stanley, 1963).
4. The clinical services, nursing units, pairs of observers, and patients were not held constant. Thus, randomly selected pairs of specialized observers were assigned to observe on randomly selected, service-specific nursing units, and randomly selected pairs of nonspecialized observers were assigned to observe on randomly selected, nonservice-specific nursing units.
  5. The Quality Monitoring Instrument, composed of various worksheets for each patient type, was not held constant between paired observations. Thus, the specific worksheet used by each pair of specialized or nonspecialized observers was determined by the randomly selected patient's classification and clinical service.
  6. Observer variables, such as status in the organization, educational level, years of nursing experience, race, age, sex, class, etc., were not controlled. Random selection and random assignment were expected to provide equivalence between groups on these and other variables.
  7. Contamination effects (e.g., discussion) may have occurred between the subjects assigned to pairs during the course of their independent observations.
  8. "Hawthorne effects" may have occurred if observers became aware of the study's purpose and hypotheses.
  9. Overlap may have occurred between the researcher's definition of specialized and nonspecialized observers.

10. The hospital's current method of assigning observers only to areas of current clinical experience may have affected the results.

Delimitations. The following delimitations have been identified:

1. The study was conducted in only one hospital. Furthermore, this hospital is a tertiary-care, teaching hospital and thus it represents a small portion of the total hospitals in Canada.
2. The study examined variables associated with the interrater reliability of only one specific quality monitoring instrument. The relationship to other quality monitoring instruments has not been established (Ventura et al. 1976).
3. The study examined only two factors associated with the interrater reliability of a quality monitoring instrument. The relationship of other factors was not examined.

### Organization of the thesis

This chapter has provided a brief introduction to the study, problem statement, definitions of pertinent terms, significance of the study, description of the variables and the hypotheses, and identification of the study's limitations and delimitations.

Chapter II contains a review of the pertinent literature related to the reliability of measurement instruments, factors affecting measurement error, and methods of assessing reliability in general, and interrater reliability in particular.

Chapter III describes the research methodology, data collection methods, and statistical treatment of the data.

Chapter IV presents the results of the data collection and the findings of the data analyses procedures.

Chapter V provides a summary of the study, followed by a discussion of the findings, implications for practice, and conclusions with recommendations for further research.

## CHAPTER II

### LITERATURE REVIEW

The literature reviewed for this experiment on variables associated with the reliability of a quality monitoring instrument focuses on four topics: reliability of measurement instruments, factors affecting measurement error, methods of assessing reliability, and assessment of interrater reliability.

#### Reliability of Measurement Instruments

Fox (1982) identified reliability as the basic attribute every data collecting instrument must possess, and as such it is a necessary precondition for validity of a measuring instrument. Kerlinger (1973) defined reliability as ". . . the accuracy or precision of a measurement instrument" (p. 443). In another approach, Polit and Hungler (1978) defined the reliability of an instrument as ". . . the degree of consistency with which it measures the attribute it is supposed to be measuring" (p. 424). Inherent in these two definitions of reliability is the concept of measurement error. Kidder (1981) noted that classical measurement theory assumes that all measurement contains some error. Accordingly, any actual score or response is composed of two components: a true component and an error component. Horn (1980) suggested that the purpose of establishing reliable measures is to reduce measurement error.

#### Factors Affecting Measurement Error

Sudman and Bradburn (1974) suggested that there are a multitude of factors that might affect the size and direction of measurement error (response error). For data collection instruments which are

directly administered to the respondents, these factors relate to the instrument and the respondent. However, for data collection instruments which require the use of raters for their administration, the numbers of factors that might affect response error are increased; not only are those factors related to the instrument and the respondent of concern but, additionally, numerous factors related to the rater, observer or interviewer must also be considered (Boyd and Westfall, 1970; Kahn and Cannell, 1963; Sudman and Bradburn, 1974).

Related to the instrument, some of the factors that should be considered are:

1. the data collection method(s) employed;
2. the construction, format, and content of questions;
3. the length and location of the interview; and,
4. the rating scale devised.

In relation to the rater, some of the factors that should be considered are:

1. the personal characteristics of the rater;
2. the selection and training process developed;
3. the interviewing skills of the interviewer;
4. the accuracy and completeness of rater recording; and,
5. the potential of rater bias and halo effect.

Related to the respondent, some of the factors that ought to be considered are:

1. the respondent's personal characteristics;
2. the respondent's perception of the question's content;
3. the respondent's motivation; and,

4. the respondent's perception of the interviewer (Kahn and Cannell, 1963; Kerlinger, 1973; Kidder, 1981; Sudman and Bradburn, 1974).

Kerlinger (1973) and Giovannetti (1981) noted that some of these factors (e.g., respondent's personal characteristics) may produce systematic or constant measurement errors, while others (e.g., rater bias) may produce unsystematic or random measurement errors. Kerlinger (1973) suggests that measurement errors are primarily random errors.

Kahn and Cannell (1963, p. 194), in an attempt to identify and classify those specific factors that relate to a data collection instrument which requires the use of interviewers, developed a model of bias in the interview. Their model identifies background characteristics, psychological factors, and behavioral factors for both the interviewer and the respondent, and can be used to facilitate investigation of the variables affecting response error. In a similar, but more comprehensive paradigm, Sudman and Bradburn (1974, p. 17) developed a model of the interview process. Their model identifies sources of response error which include the respondent, the interviewer, and the instrument. In addition, they coded 46 independent variables that may affect response error into three categories: task variables, interviewer role, and respondent role (Sudman and Bradburn, 1974, p. 21).

In the social sciences literature, Sudman and Bradburn (1974) noted that hundreds of studies on many of these factors have been reported. They cautioned, however, that many of these studies were conducted in highly specific situations from which it is difficult

to generalize. Furthermore, they suggested that the investigation of these factors had been complicated by the fact that much of the research was conducted without any theoretical framework and therefore the results did little to advance an understanding of the ways the factors affect response error. Nevertheless, from their review and synthesis of the existing research on response effects, Sudman and Bradburn (1974) reported that research studies have found, in regard to task variables, that:

1. whether the question was closed or open ended appears to cause minimal response error;
2. the method of questionnaire administration, self versus interview, was a more important variable in response error than location of the interview;
3. the differences between self-administered and interview questionnaires were larger than differences due to respondent and interview characteristics;
4. threatening and non-threatening behavioral and attitudinal questions produced minimal response error; and,
5. non-salient questions increased response error.

In regard to the role of the interviewer, Sudman and Bradburn (1974) reported that research studies had shown that:

1. response error was twice as high for inexperienced interviewers;
2. higher social status interviewers induced a larger response error;
3. response error declined in interview situations as the age of the interviewer increased; and,
4. open ended questions were influenced by the sex of the interviewer.



As well, Steinkamp (1966) reported that an effective interviewer scored higher on the Edwards Personal Preference Schedule in areas of dominance, self-confidence, and attention to detail.

In regard to the respondent's role behavior, Sudman and Bradburn (1974) reported that research studies found that:

1. differences in sex, race, and age of the respondents did not influence response error;
2. the percent of "don't know" responses declined as education increased;
3. the percent of "don't know" responses was higher among females; and,
4. a larger response error was found for females in threatening close ended questions where a socially desirable response is available.

Sudman and Bradburn (1974) concluded that whereas the role of the interviewer had received a lion's share of research on effect and control because interviewer effects are such an obvious potential source of bias, the "... nature of the task and the conditions under which it is performed are among the variables that have the strongest effects on response. These variables are typically far larger than the effects due to interviewer characteristics" (p. 28). The above findings have potential relevance for the reliability of quality monitoring instruments in health care organizations, particularly those which require the use of an interviewer.

In health services research, factors affecting the reliability of data collecting instruments have received minimal attention.

Regarding observers, Hegyvary and Haussmann (1976) reported that one of the greatest problems in quality monitoring is that of observer reliability. They noted the need for the investigation of variables associated with reliability such as observer education, personal characteristics, and position in the organization. Similarly, Ventura et al. (1980) reported that although the selection of observers is an important activity, the topic had received minimal attention in the literature. One notable exception was a study by Haussmann, Hegyvary, and Newman (1976, p. 21) which reported that personal characteristics of observers as measured by the California Psychological Inventory (CPI) and the Watson-Glaser Inventory did not provide an adequate basis for observer selection. They concluded that extensive training sessions and reliability testing at least every month by observers were the significant factors in the appropriate and reliable use of the Rush-Medicus Instrument and that the personal characteristics of observers were less important in ensuring the quality of the data. In respect to data collection methods, the literature is equally sparse. One notable exception was a study by Herman and Cayten (1980) which found low interrater reliability on medical record abstraction was associated with variables which required the rater to use judgement.

The prevalence today of ongoing data collecting instruments in health care organizations, combined with the minimal attention in the literature with respect to factors affecting the reliability of these instruments, supports the need for research in this area. For instruments which require the use of raters, research on factors affecting interrater reliability is needed. For instruments of this

nature, Ventura et al. (1980) noted that interrater reliability was the most critical reliability assessment method.

### Methods of Assessing Reliability

Polit and Hungler (1978, p. 426) stated that the reliability of a measuring instrument could be assessed by several different methods. The particular method chosen depended on the nature of the instrument and on the aspect of the reliability concept that was of greatest concern. They noted that three aspects (types) of reliability had received major quantitative attention: stability, internal consistency, and equivalence. Goodwin and Prescott (1981) noted that each of these common types of reliability differs in its operational definition of consistency and in its operational method. For example, stability or test-retest reliability refers to the consistency of scores obtained on repeated administrations of the instrument; internal consistency reliability refers to the consistency of the individual's responses to various subsets of items comprising the instrument; and, equivalence by alternate forms refers to the consistency of scores obtained by more than one form of an instrument, whereas equivalence by interrater reliability refers to the consistency of different raters, observers, interviewers, or judges in their use of the instrument (Goodwin and Prescott, 1981; Polit and Hungler, 1978). Goodwin and Prescott (1981) noted that regardless of which of the different aspects of reliability is of interest ". . . the basic theoretical meaning of a reliability coefficient is the same: the amount of variance in a set of scores that is not the result of errors of measurement"

(p. 324). Furthermore, they suggested that for many instruments, more than one aspect of reliability should be assessed.

### Assessment of Interrater Reliability

For data collection instruments, which require the subjective opinions of raters, scorers, observers, interviewers or judges, Goodwin and Prescott (1981) noted that interrater reliability is the preferred method of assessing reliability. Interrater reliability is an estimate of the degree to which two or more independent raters, observers, scorers, judges, or interviewers are consistent in their judgements (Goodwin and Prescott, 1981, p. 325). These researchers suggested that there are several approaches to assessing interrater reliability: percentage of agreement, correlational techniques, comparison of means, and generalizability theory techniques (an extension of the intraclass correlation technique). The percentage of agreement approach expresses reliability as the number of times the raters or judges agree relative to the total number of observations made. The correlational techniques express reliability in terms of correlations between the sets of scores of two raters; that is, the extent to which events or subjects are ranked similarly by the different raters or observers. Comparisons of means expresses reliability in terms of agreement between actual scores of subjects and it is calculated using t-tests or analysis of variance (ANOVA). (Anova can also be used to obtain intraclass correlation coefficients as estimates of reliability.) Finally, the generalizability theory approach encompasses all classical reliability approaches and provides a comprehensive method for assessing numerous

potential influences on measurement error (Goodwin and Prescott, 1981, p. 325).

The appropriate interrater approach to use in estimating the consistency of a measuring instrument is itself subject to inconsistency in the literature. On the one hand, for example, Goodwin and Prescott (1981, p. 325) noted that when one is working with categorical, nominal data the percentage of agreement estimate is both an appropriate and sufficient approach to interrater reliability. They further noted, that this approach is especially appropriate when the data have a narrow range, "the larger the number of choices available to raters (or the finer the distinctions possible), the higher the probability that exact agreement will not occur" (Goodwin and Prescott, 1981 p. 330). Similarly, Maguire and Hazlett (1969, p. 125) also noted that when one is working with nominal data, percentage of agreement is appropriate as the estimate of consistency. They stated that numerical methods, such as correlation techniques and comparisons of means, depend upon interval data being available. In addition, the correlation techniques require that the study design include fixed pairs of raters. They further observed that even when one is working with interval data, it may be useful to look at the percentage agreement statistic to estimate consistency. Moreover, they observed, that the percentage agreement statistic has the desirable property of being understood by a person who is statistically naive (Maguire and Hazlett, 1969, p. 125).

On the other hand, Bartko and Carpenter (1976) argued that simple percentage of agreement ignores chance agreements which can be numerous

if few categories are used by the raters. Instead, for dichotomous, nominal scale data utilizing two raters they recommended the Kappa statistic to correct for chance agreement, and for dichotomous, nominal scale data utilizing more than two raters they recommended the intraclass correlation coefficient. (Again the requirement of a study design with fixed pairs is needed.) In the case of polychotomous, nominal data utilizing two raters they recommended the Kappa and weighted K; whereas, for polychotomous, nominal data utilizing more than two raters they recommended the generalized Kappa. For quantitative data they recommended an analysis of variance (ANOVA) intraclass correlation approach. Similarly, for quality monitoring instruments which require the use of raters, Ventura et al. (1980) contended that the intraclass correlation coefficient (ICC) is a more appropriate method than percentage agreement for assessing rater agreement.

Notwithstanding the above inconsistencies, the approach and formula for estimating interrater reliability depends upon the nature of the instrument, the type of data, the research design, and the use of the measurement. As Maguire and Hazlett (1969) observed:

the question of reliability of a measure is fundamentally a question of the consistency of the measurement. There are many ways to calculate indices of consistency, the one that is used should be determined by the use to which the measurement is put, and not by blind obedience to common practice (p. 125).

## CHAPTER III

### METHODOLOGY

#### Type of Study

This study, on the interrater reliability of a quality monitoring instrument, was a field experiment utilizing the Posttest - Only Control Group design (Campbell and Stanley, 1963). This design takes the following form:

$$\begin{array}{ccc} R & x & , O_1 \\ R & & O_2 \end{array}$$

The advantages of this design are as follows:

1. It is less expensive and time consuming to implement than designs which include the use of a pretest.
2. The absence of a pretest reduces the potential for observer-interaction effects occurring among the subjects assigned to pairs.
3. The random assignment of subjects to groups (experimental and control) ensures equivalence of the two groups.
4. The random assignment of subjects to groups, plus the absence of a pretest, will control for factors threatening internal validity, such as history, maturation, testing, instrumentation, regression, and selection biases.
5. Any differential loss of subjects which may occur (i.e., experimental mortality) can be considered in the data analysis procedures (Campbell and Stanley, 1963).

#### Setting

A large tertiary-care teaching hospital served as the setting

for this study. This hospital has 994 acute care beds and 320 long term care beds. The hospital is affiliated with a university nursing and medical school. In addition, the study hospital is affiliated with a diploma nursing school.

A random selection of nursing units on the various clinical services served as data collection sites.

In 1979, the Nursing Department of this hospital implemented the Rush-Medicus Quality Monitoring Instrument. Each month, approximately one-third of the 52 inpatient nursing units are scheduled for quality monitoring observations (Appendix C).

Quality on any nursing unit is monitored on the basis of a review of 10 percent of that nursing unit's patient admissions. The quality observations for a nursing unit are distributed randomly across the month so that 60 percent of the observations occur on days and 40 percent occur on evenings and weekends. No observations are scheduled for the night shift. A single observation (commonly referred to as an audit) consists of the assessment of two randomly selected patients plus a general unit assessment. At this hospital, an average of 115 observations are scheduled for each month. Consequently, an average of 230 patients and 115 units are assessed each month. A master schedule defines the number and dates of observations required for each nursing unit. Trained nurse observers "sign up" for the observations that they will conduct. Prior to the start of an actual observation, the nurse observer randomly selects two patients and determines their illness classification. According to the patient's classification level, appropriate worksheets (questionnaires) are



selected as well as the specific questionnaire for the unit observation. One observation requires an average of 1½ - 2 hours to complete (Medicus, undated).

### Population of Specialized Observers

The majority of nurse observers (170 out of 183 as of December 1983) at this hospital were decentralized to their own clinical service areas for the purpose of conducting quality monitoring observations. Decentralized nurse observers conduct observations only within their respective clinical service areas, with the exception of their own nursing unit. For example, nurse observers from medical units conduct observations only on medical units. For this study, these decentralized nurse observers were defined as specialized observers and thereby constituted the study population.

Specialized nurse observers are selected by either their position in the organization or through volunteering: unit supervisors (charge nurses), clinical instructors, and assistant directors of nursing are expected to conduct observations as part of their position's responsibilities; on the other hand, only those general duty nurses who volunteer to do so become observers. Accordingly, this population of nurse observers represents nurses with varying educational backgrounds, differing organizational status, varying clinical experience, and varying observer experience.

All nurse observers receive a standard training program of 8 - 10 hours in length, prior to becoming observers. This observer orientation program includes preparatory reading material, classroom instruction, a trial observation with an experienced observer, and

an interrater reliability test. No formal system currently exists for either regular retraining sessions or regular interrater reliability testing subsequent to the initial orientation program. However, individual and group problem solving sessions are provided on an ad-hoc basis.

### Sampling Plan

The population of 170 specialized nurse observers was stratified into seven subgroups representing the existing clinical service areas. These subgroups are medicine, surgery, pediatrics, obstetrics, long term care, psychiatry, and special care areas. The numbers of specialized observers within each subgroup were as follows:

medicine	37
surgery	63
obstetrics	22
long term care	11
pediatrics	07
psychiatry	14
special care	16

The above stratification was required for this study in order to assign the control group (specialized observers) to clinical areas within their clinical subgroup and to assign the experimental group (nonspecialized observers) to clinical areas outside their clinical subgroup.

### Random Sampling of Subjects

A simple random selection of 62 observers (subjects) was drawn from within the various subgroups. For each subgroup, a goal of

6 - 12 randomly selected subjects was set. The number of subjects randomly drawn from within each subgroup was as follows:

medicine	12
surgery	12
obstetrics	08
long term care	08
pediatrics	06
psychiatry	08
special care	08

Random sampling was accomplished by using a random numbers table. This is a procedure to select subjects to study that ensures the sample is free from selection bias (Stuart, 1957).

#### Random Assignment of Subjects to Pairs

Following the random selection of observers from each subgroup, the selected observers within each subgroup were assigned to pairs by the hat-draw method. Kidder (1983) notes that "Provided you have the names all written on similar slips of paper and have shuffled the slips sufficiently, you have a random assignment procedure as good as any" (p. 18). This random assignment procedure yielded a total of 31 pairs. The number of pairs for each subgroup was as follows:

medicine	06
surgery	06
obstetrics	04
long term care	04
pediatrics	03

psychiatry	04
special care areas	04

#### Random Assignment of Pairs to Experimental or Control Group

For each subgroup, each pair of observers was randomly assigned to either experimental or control group status by the hat-draw method. This random assignment process yielded 15 experimental pairs and 16 control pairs. It also yielded, on two selected variables, the following equivalence between the two groups:

1. On the variable, status in the organization, the control group consisted of 13 general duty nurses, 11 unit supervisors, 6 clinical instructors, and 2 assistant charge nurses, whereas the experimental group consisted of 13 general duty nurses, 9 unit supervisors, and 8 clinical instructors.
2. On the variable, years of experience as a nurse observer, the control group's mean was 2.5 years whereas the experimental group's mean was 2.8 years.

#### Random Assignment of Pairs to Nursing Units

Each experimental pair was assigned to conduct an interrater observation on a randomly selected nursing unit outside their respective subgroup. Similarly, each control pair was assigned to conduct an interrater observation on a randomly selected nursing unit within their subgroup (except their own unit). This random assignment procedure was accomplished by the hat-draw method. This random assignment of 15 experimental and 16 control pairs to nursing units yielded the following equivalence of assigned clinical service areas between the two groups:

<u>Clinical Service</u>	<u>Experimental (pairs)</u>	<u>Control (pairs)</u>
medicine	2	3
surgery	2	3
obstetrics	1	2
long term care	2	2
pediatrics	4	2
psychiatry	2	2
special care	2	2

### Instrumentation

The Rush-Medicus Quality Monitoring Instrument was used by the pairs of nurse observers in both the experimental and control group. This instrument consists of 440 questions related to 357 criteria. The criteria are grouped into homogeneous clusters to assess the following six objectives of the nursing process:

1. The plan of nursing care is formulated.
2. The physical needs of the patient are attended.
3. The nonphysical needs of the patient are attended.
4. The achievement of objectives is evaluated.
5. The unit procedures are followed for patient protection.
6. The delivery of nursing care is facilitated by administrative and managerial services.

These six objectives are further delineated into 28 subobjectives (Appendix D). A representative sample of each subobjective's eligible pool of criteria is randomly assigned to a series of questionnaires relevant to patient type (as determined by patient classification)

and to clinical service (Jelinek, Hausmann, Hegyvary, and Newman, 1975) (Appendix E).

For the purpose of this study, each questionnaire, composed of a random selection of questions from the total Rush-Medicus Quality Monitoring Instrument, was considered equivalent and therefore was not held constant. Because each questionnaire contains only a relatively small random sample (i.e., 30-40) of the total questions, the assessment of the quality of nursing care on any single nursing unit requires a minimum of six complete observations (i.e., 12 patients and 6 unit questionnaires) in order to achieve a reliable score (Medicus, undated).

#### Data Collection Methods

A source of information is specified for each of the 440 questions in the Rush-Medicus Instrument. There are eight different sources from which the nurse observer collects data on each observation. These include patient record, patient observation, patient interview, nursing personnel interview, nursing personnel observation, patient environment observation, observer inference, and unit management observation (Appendix F). Accordingly, a nurse observer using the Rush-Medicus instrument collects data by four different methods: record abstraction, interview, observation, and observer judgement.

#### Measurement of the Rush-Medicus Instrument

The majority of the 440 questions (89 percent) contained in the Rush-Medicus Instrument require the observer to select a nominal, categorical, dichotomous response. A minority of the 440 questions (11 percent) require the observer to select an ordinal response.

That is, 391 questions require the observer to select from and code either a "yes" or "no" response, whereas 49 questions require the observer to select from varying degrees of "yes," such as; yes, complete or incomplete; yes, some of the time, most of the time, or all of the time; in addition to the "no" category. For all the questions, the observer is also directed to code "not applicable" or "information not available" whenever appropriate. This latter coding in effect removes that particular question from the quality assessment score.

Quality scores for a particular nursing unit are expressed as the ratio of positive responses to the maximum possible positive responses after those questions which were not applicable have been excluded. All responses are treated equally; that is, no attempt was made to weight them in terms of their relative importance or contribution to the particular attribute of nursing that is being addressed by that subobjective (Hausmann, Hegyvary, and Newman, 1976, p. 11). Scores, expressed in percentage, are computer calculated for each subobjective, each objective, and a total score based on the average of the first four objectives. As mentioned previously, a minimum of six complete observations is considered necessary in order to provide reliable scores for any nursing unit on the subobjectives, objectives, and total score.

#### **Reliability of the Rush-Medicus Instrument**

Hegyvary and Hausmann (1976) report that during the initial testing period of the Rush-Medicus Instrument, the interrater reliability, using the percentage agreement approach, ranged between 83

and 92 percent. This range, however, was founded on interrater reliability estimates for only six patients (three in each of the two testing hospitals).

The consulting firm that distributes this particular instrument to hospitals, recommends a minimum goal of 85 percent agreement for the interrater reliability of this instrument. As mentioned previously, the study hospital did not formally test for interrater reliability beyond the orientation period. However, interrater reliability tests (62 tests over a one year period) associated with the orientation of new observers found percentage agreements that ranged from 68 to 97 percent, with a mean of 87.4 percent.

In the previously reported survey of Canadian hospitals using the Rush-Medicus Instrument, the ten hospitals that conduct regular interrater reliability testing reported percentage agreements that ranged from 85 to 98 percent, with a mean of 90 percent and a mode of 95 percent.

Ventura et al. (1980) argue that the percentage agreement statistic is inappropriate for estimating the reliability of instruments of this nature. Instead they recommend the Intraclass Correlation Coefficient (I.C.C.) because it accounts for differences among raters in level, as well as differences in ranking. In their research on the interrater reliabilities for two measures of nursing care quality (i.e., the Rush-Medicus Instrument and the QualPacs Instrument) they found that the Rush-Medicus Instrument met the standard criterion of .75 in only 12 of the 20 instances of testing. However, it is noted that Ventura et al. (1980) used the actual quality scores for



each patient in determining the I.C.C. value. In addition, they excluded the unit observation. As previously mentioned, the Rush-Medicus Instrument does not purport to provide a reliable score based on less than six observations (i.e., 12 patients and 6 unit observations). For example, the number of responses that would be available from one patient are so minimal that percentage scores for each subobjective could be based on as few as one question per subobjective (e.g., if yes = 100 percent, if no = 0 percent). Furthermore, their exclusion of the unit observation reduced even further the questions applicable to objective 5 and objective 6.

#### Validity of the Rush-Medicus Instrument

Validity refers to the ability of an instrument to measure what it is intended to measure. Jelinek et al. (1974) report that extensive statistical analyses were carried out on the Rush-Medicus Instrument's initial data to evaluate the worth of the criteria as measures of quality. They report that the methodology has proven validity. However, like reliability, validity is not necessarily constant from one setting to another. At the study hospital, face validity of the Rush-Medicus Instrument is established annually by a panel of nurse experts. Each criterion is assessed for clarity, appropriateness to the study hospital, and relevance to the objective it is supposed to measure. In addition, the Rush-Medicus Instrument was assessed for relevance to the Nursing Standards developed by the Alberta Association of Registered Nurses (AARN) and the Canadian Nurses Association (CNA). In both instances, the panel of experts agreed that the criteria reflected the above standards. For the purpose

of this study on reliability, the validity of the Rush-Medicus Instrument was accepted, but it was not a central concern of this study. Fox (1982) contends that only if an instrument is reliable do we worry about whether it has the other characteristics (p. 255). He contends that validity is the second most important characteristic and one for which reliability is a precondition (Fox, 1982 p. 260).

### Study Design

Over a three month period (January - March, 1984), each of the 16 control group pairs (specialized observers) was scheduled to conduct one interrater reliability observation, during a certain month, on a randomly selected nursing unit within their area of specialization. Similarly, over the same period, each of the 15 experimental group pairs (nonspecialized observers) was scheduled to conduct one interrater reliability observation, during a certain month, on a randomly selected nursing unit outside their area of specialization. (For the control group, this meant a total of 32 patient and 16 nursing unit assessments, and for the experimental group a total of 30 patient and 15 nursing unit assessments.) (Appendix G).

Prior to the commencement of the study, each of the randomly selected subjects was contacted. At this time, each subject was given a brief description of the study and provided the opportunity to participate. Three of the randomly selected subjects, one each from medicine, pediatrics, and psychiatry were unable to participate because of either general illness, plans to terminate, or permanent night shift. Random replacement of the above three from each respective subgroup was accomplished. A research consent form was

obtained from each subject (Appendix H). A total of 62 subjects agreed to participate. Once agreement to participate was obtained, the subjects were given a verbal and a written explanation of interrater reliability testing (Appendix I). They were also informed of their partners and of the month in which they were scheduled to conduct the observation. As well, the control pairs were informed of the nursing unit to which they had been randomly assigned within their area of specialization. The experimental pairs were only informed that they may be required to conduct an interrater observation on any nursing unit within the hospital (the specific unit was indicated immediately prior to the actual observation). All pairs were requested to arrange the day, and time of day, to conduct the observation among themselves, at their convenience, and notify the investigator of their decision.

### **Ethical Considerations**

In this study, the anonymity of the subjects and patients was protected. For the patients, the study was conducted as part of the existing quality assurance program. The existing program requires that each nurse observer request the randomly selected patient's agreement to participate while assuring the patient of anonymity and confidentiality. The above practice continued during the study and was considered sufficient because the study's findings relate only to the interrater agreement of the observers and not to the actual quality scores or comments of the patients. For the observers, as mentioned previously, each nurse observer's participation was voluntary and a research consent form was signed by each subject.

### Data Collection Procedure

On the day or evening chosen by either a control or experimental pair, the observers obtained the customary audit pack which contained duplicate copies of the questionnaires and answer sheets from the Audit Office (Appendix J). At this time, the experimental pairs were informed of the nursing unit where they were expected to conduct their observation. The procedure for interrater reliability testing was again briefly reviewed at this time with emphasis on a coin-toss to determine which of the two should interview the first patient. On the unit, each pair conducted the patient and unit observations according to the established procedure for quality monitoring and interrater reliability testing. The average interrater observation required approximately 2 hours for the control pairs and 2½ hours for the experimental pairs. Following completion of the observation of two patients and the unit, each pair returned to the investigator's office for calculation of percentage agreement and a debriefing on each question and response in disagreement. The percentage agreement score and each observer's comments related to the responses in disagreement, were recorded for each pair (Appendix K). The debriefing sessions lasted an average of 1 hour. One observer's answer sheet became the official observation, while the other answer sheet, with identification of the specific questions in disagreement, was filed for the data analyses required for this study.

During the first month of the study, an instance of contamination during observations became apparent for an experimental pair. This pair, during debriefing, indicated that they had discussed a few

responses related to the patient's chart. Thus, their percentage agreement score was eliminated and another interrater observation required. Consequently, for the remaining months of the study, all pairs were provided with a verbal comprehensive review of the interrater reliability procedure, by the investigator, prior to the commencement of their observation. This review emphasized the necessity for independent but simultaneous observations in order for the interrater observation to be of any value. As well, the need to independently, without discussion, review each patient's chart was stressed. No further instances of contamination were apparent.

The study extended into the month of April because of the difficulty some of the pairs, particularly general duty nurses on varying shifts and/or very busy nursing units, were experiencing in finding a mutually agreeable time to conduct the interrater observation. One control pair of general duty observers were eventually excluded from the study because they were unable to find a mutually convenient time.

### Data Analyses

A special computer program, separate from the one provided for the Rush-Medicus Quality Monitoring Instrument, was designed to accommodate the analyses required by this study.<sup>1</sup> This program included the Rush-Medicus Master list of criteria and each

---

<sup>1</sup> This special program was designed and operated by C. Prokop, Department of Educational Administration, University of Alberta.

questionnaire's list of criteria and related questions. Each criterion was numerically identified by source of information, objective, and subobjective.

Following the completion of the data collection process, the control and experimental pairs' answer sheets, which identified the specific questionnaire used as well as the specific questions in disagreement, were introduced into the computer program for the analyses chosen for this study: percentage agreement and Chi-square calculations.

The simple percentage of agreement calculation was chosen and considered appropriate and sufficient for this study, rather than the more sophisticated Kappa or ICC statistic, for the following seven reasons:

1. The Rush-Medicus Instrument provides basically nominal dichotomous data (e.g., of the instrument's 440 questions, 391 questions require nominal, dichotomous responses and only 49 questions require ordinal, polychotomous responses).
2. The chance agreement on dichotomous, nominal scales is obviously 50 percent.
3. The study design neither included fixed pairs nor patients held constant between pairs.
4. The Rush-Medicus Methodology used percentage agreement among observers.
5. The actual scores per patient are considered unreliable.
6. The study's results will be compared with other Rush-Medicus users who use the percentage agreement approach.

7. The study's results will be shared with nursing administrators and others who are more familiar with the meaning of the percentage agreement statistic.

Accordingly, for both the experimental and control groups, a percentage of agreement statistic was calculated for each objective, as well as for all objectives combined. In addition, for both the experimental and control groups, the percentage of agreement statistic was calculated for each source of information utilized by the Rush-Medicus Instrument. A reliability of .85, similar to the goal recommended in the Rush-Medicus Methodology, was considered acceptable. The above calculations of percentage agreement were based on assessments of 30 patients and 15 unit observations for each group. The random selection of patients by each pair yielded the following equivalence of questionnaires, by patient type, used by each group:

Questionnaire by Patient Type	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
Experimental Group	9	13	8	0	(30)
Control Group	13	8	6	3	(30)

The Chi-square statistic, two variable case, was considered the appropriate test of significance because the data were essentially nominal. The Chi-square tests the goodness of fit of the observed to expected distributions and finds the differences between the observed and expected distributions per cell (Broyles and Lay, 1979). The Chi-square statistic, set at .05 level of significance, was used to determine the significance, if any, of the observed differences in numbers of disagreements (proportions) between the two groups on all objectives, each objective, and between and within the two

groups on sources of information. As well, the t-test was used to determine the difference, if any, between the means of the percentage agreement scores obtained by the experimental and control pairs.

Accordingly, the first null hypothesis was tested using both the Chi-square analyses and the t-test of differences between means of independent samples. The second and third hypotheses were tested using the Chi-square analyses. The Yates correction was used, whenever appropriate, when the expected cells were less than ten (Ferguson, 1966, p. 207).

In addition, following completion of the data collection period, each observer's comments, recorded during the debriefing sessions, relating to questions found to be in disagreement between the pair were coded by the investigator, with the assistance of a research assistant, into one of the following nine categories representing reasons for disagreement:

1. Error in recording
2. Patient response unclear
3. Nurse response unclear
4. Question unclear
5. Hospital standard on policy or procedure varies
6. Observer judgement varies
7. Careless reading of question
8. Patient record confusing
9. Other reasons.

For each of the above categories, a few representative examples



of the observer's comments that were coded into each category are as listed below:

1. Error in Recording - included comments such as;

"I meant to circle 1 not 2,"

"I circled the wrong response,"

"I forgot to circle a response."

2. Patient's Response Unclear - included comments such as;

"I couldn't understand the patient's answer re: leg exercises,"

"I felt the patient's response was ambiguous,"

"I took the patient's nod of his head as a 'yes' response,"

"The patient was crying, I couldn't understand his answers,"

"I think the patient was confused."

3. Nurse's Response Unclear - included comments such as;

"I misunderstood the nurse's response,"

"The nurse's answer was unclear,"

"The nurse said no chance (one observer recorded no and the other recorded not applicable),"

"The nurse's answer was ambiguous."

4. Question Unclear - included comments such as;

"The question is not specific enough,"

"The question is unclear,"

"The question is double-barrelled."

5. Standard Varies - included comments such as;

"The hospital standard varies regarding procedures for cardiac arrest,"

"The hospital standard is unclear regarding actions to be taken during a fire,"

"The hospital standard varies regarding order of the chart,"

"The hospital policies were unclear regarding nurses notes,"

"The hospital charting guidelines are unclear regarding abbreviations."

6. Judgement Varies - included comments such as;

"I used my personal judgement regarding appropriateness of discharge teaching,"

"I used my personal judgement regarding garbage can sufficiently empty,"

"Judgement varied regarding what constitutes current treatment,"

"Judgement varied regarding presence of tubes,"

"Judgement varied regarding what constitutes noise and corridor clear."

7. Careless Reading of the Question - included comments such as;

"I didn't notice the time frame,"

"I didn't read the source of information,"

"I didn't notice response options on next page,"

"I didn't read difference between some of the time and all the time,"

"I didn't read section relating to "not applicable."

8. Record Confusing (included careless reading of the patient's record) - included comments such as;

"I missed the charting on the back of the form,"

"I felt the record was unclear,"

"Routine orders were confusing,"

"Records unclear related to discharge teaching,"

"I didn't notice the treatment orders written,"

"I found the chart confusing,"

"I didn't carefully read the record."

9. Other - included comments such as;

"I can't recall why I circled 'yes' ,"

"Observer's saw different situations,"

"Can't recall."

The above coding frame and comments were not subjected to formal reliability testing procedures (Kidder, 1978, p. 305).

Following the coding of each comment into one of the above nine categories, each of the questions in disagreement on each pair's original answer sheets was subsequently coded according to one of the above reasons. An analysis of the reasons for disagreements between and within the experimental and control groups was performed. This analysis was calculated using percent differences and also the Chi-square statistic. The Yates correction was used, whenever appropriate, when the expected cells were less than ten.

Furthermore, for those specific objectives in which there were significant differences noted between the experimental and control groups in terms of numbers of disagreements, the reasons for these disagreements were analysed for possible significant differences.

Finally, a computerized list was produced of each question that was used during the study by both the control and experimental pairs. This list identified the frequency of disagreement and the reasons

for disagreement related to each question. Although this list was not needed for the analyses required by this study, it is invaluable for future orientation and retraining sessions of nurse observers at the study hospital (Appendix L).

In summary, the analyses for this study included:

1. Percentage of agreement, all observations combined, for each group.
2. Percentage of agreement, by objective, for each group.
3. Percentage of agreement, by source of information, for each group.
4. Chi-square analyses, between group's, on total disagreements.
5. T-test of differences between groups' mean scores.
6. Chi-square analyses of disagreements, between groups, by objectives.
7. Chi-square analyses of disagreements, between groups, by source of information.
8. Chi-square analyses of disagreements, within each group, by source of information.
9. Reasons per group estimated as percent of total disagreements.
10. Chi-square analyses of disagreements, between groups, by reasons.
11. Chi-square analyses of disagreements, within groups, by reasons.
12. Chi-square analyses of disagreements, between groups, by subobjectives relative to a significant objective.
13. Chi-square analyses of disagreements, between groups, by source of information relative to a significant objective.
14. Chi-square analyses of disagreements, between groups, by reasons relative to a significant objective.

## CHAPTER IV

### RESULTS

The findings of this field experiment, on variables associated with the interrater reliability of a quality monitoring instrument, are presented in the following order: interrater reliabilities estimated by percentage agreement, significant differences and hypotheses testing, reasons for disagreements, and analyses of significant objectives.

#### Interrater Reliabilities Estimated by Percentage Agreement

A minimum of 85 percent agreement is recommended for the interrater reliability of the Rush-Medicus Instrument (Medicus, undated).

Percentage agreement by observation. As shown in Table 1, the average percentage agreement for all observations conducted by the experimental group exceeded the minimum acceptable level of 85 percent agreement. The average percentage agreement for all observations conducted by the control group did not achieve the minimum acceptable level of 85 percent agreement. The experimental group's mean score, based on 15 paired observations ( $n = 1651$ ), was 87.2 percent, with a range of 77 - 98 percent,  $Md = 87$  percent, and  $Sd = 5.5$ . Of these fifteen paired observations, four experimental pairs achieved less than 85 percent agreement. The control group's mean score, based on 15 paired observations ( $n = 1657$ ), was 82.8 percent, with a range of 77 - 95 percent,  $Md = 86$  percent, and  $Sd = 5.1$ . Of these fifteen paired observations, nine control pairs achieved less than 85 percent agreement.

**TABLE 1**  
**INTERRATER RELIABILITY FOR EACH OBSERVATION**  
**ESTIMATED BY PERCENT AGREEMENT**

<sup>a</sup> Observation Number	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
<sup>b</sup> Experimental Pairs Percent	87	77	98	87	89	86	82	96	90	86	85	90	80	90	84
<sup>c</sup> Control Pairs Percent	80	78	86	89	87	80	77	78	82	85	82	86	78	79	95

a = one observation included two patients and a unit assessment.

b = experimental group,  $\bar{x}$  = 87.2, Md = 87, Range 77 - 98, Sd = 5.5.

c = control group,  $\bar{x}$  = 82.8, Md = 86, Range 77 - 95, Sd = 5.1.

**Percentage agreement by objective.** As shown in Table 2, of the six objectives measured by the Rush-Medicus Instruments' criteria, the experimental group achieved or exceeded the acceptable level of 85 percent agreement on criteria related to four objectives:

Objective 2, Physical Needs Attended (91.7 percent)

Objective 3, Nonphysical Needs Attended (88.9 percent)

Objective 5, Unit Procedures Followed (88.7 percent)

Objective 6, Delivery of Nursing Care Facilitated (92.6 percent)

The control group achieved or exceeded the recommended level on criteria related to two of the six objectives:

Objective 3, Nonphysical Needs Attended (85.6 percent)

Objective 5, Delivery of Nursing Care Facilitated (90.0 percent)

**TABLE 2**  
**INTERRATER RELIABILITY BY OBJECTIVES**  
**ESTIMATED BY PERCENT AGREEMENT**

<u>Objectives</u>	<u><sup>a</sup>Number of Observations</u>	<u><sup>b</sup>Experimental</u>	<u><sup>c</sup>Control</u>
(1) Nursing Care Plan is Formulated	15	77.6	76.2
(2) Physical Needs are Attended	15	91.7 *	82.8
(3) Nonphysical Needs are Attended	15	88.9 *	85.6 *
(4) Nursing Care Objectives are Evaluated	15	65.6	65.6
(5) Unit Procedures Followed	15	88.7 *	83.6
(6) Delivery of Nursing Care Facilitated	15	92.6 *	90.0 *
All Objectives	15	87.2 *	82.8

**Note:** Minimum acceptable level = 85 percent agreement.

a = 15 paired observations per group.

b = experimental group = 15 pairs of nonspecialized observers.

c = control group = 15 pairs of specialized observers.

\* = minimum acceptable level achieved or exceeded.

Neither the experimental nor the control group achieved the minimum acceptable level of agreement on criteria related to Objective 1, Nursing Care Plan is Formulated, (77.6 percent and 76.2 percent respectively) and Objective 4, Nursing Care Objectives are Evaluated (65.6 percent and 65.6 percent respectively). In addition, the control group did not achieve the minimum acceptable level of agreement on criteria related to Objective 2, Physical Needs Are Attended, (82.8 percent) and Objective 5, Unit Procedures Followed, (83.6 percent).

In summary, the experimental group exceeded the acceptable level of 85 percent agreement on all the observations (15) combined. They achieved or exceeded the acceptable level of percentage agreement on four of the six objectives measured by the Rush-Medicus Instrument. The control group did not achieve the acceptable level of 85 percent agreement on all the observations (15) combined. As well, they only achieved the acceptable level of percentage agreement on two of the six objectives measured by the Rush-Medicus Instrument (Table 2).

Percentage agreement by source of information. Each of the Rush-Medicus Instrument's criterion requires the nurse observers to collect data from one of eight different sources. As shown in Table 3, of the eight sources, the experimental group achieved or exceeded the minimum acceptable level of 85 percent agreement for criteria obtained from seven sources:

Source	II, Patient Observation	(94.6 percent)
Source	III, Patient Interview	(89.2 percent)
Source	IV, Nursing Interview	(93.0 percent)



TABLE 3  
 INTERRATER RELIABILITY BY SOURCE  
 ESTIMATED BY PERCENT AGREEMENT

<u>Sources</u>	<u><sup>a</sup>Number of Observations/ Group</u>	<u><sup>b</sup>Experimental Group Percent Agreement</u>	<u><sup>c</sup>Control Group Percent Agreement</u>
I Patient Record	15	76.8	73.4
II Patient Observation	15	94.6 *	85.1 *
III Patient Interview	15	89.2 *	88.5 *
IV Nursing Interview	15	93.0 *	90.5 *
V Nursing Observation	15	92.8 *	81.1
VI Environmental Observation	15	93.2 *	84.8
VII Observer Inference	15	95.0 *	91.7 *
VIII Unit Management Observation	15	100.0 *	66.7
All Sources	15	87.2	82.8

**Note:** Minimum acceptable level = 85 percent agreement.

a = 15 paired observations per group.

b = experimental group = 15 pairs of nonspecialized observers.

c = control group = 15 pairs of specialized observers.

\* = minimum acceptable level achieved or exceeded.

Source	V,	Nursing Observation	(92.8 percent)
Source	VI,	Environmental Observation	(93.2 percent)
Source	VII,	Observer Inference	(95.0 percent)
Source	VIII,	Unit Management Observation	(100 percent)

The control group achieved or exceeded the minimum acceptable level of 85 percent agreement for criteria obtained from four of the eight sources:

Source	II,	Patient Observation	(85.1 percent)
Source	III,	Patient Interview	(88.5 percent)
Source	IV,	Nursing Interview	(90.5 percent)
Source	VII,	Observer Inference	(91.7 percent)

Neither the experimental nor the control group achieved the acceptable level of percentage agreement for criteria obtained from Source I, Patient Record, (76.8 percent and 73.4 percent respectively). In addition, the control group did not achieve the acceptable level of percentage agreement for criteria obtained from three other sources:

Source	V,	Nursing Observation	(81.1 percent)
Source	VI,	Environmental Observation	(84.8 percent)
Source	VIII,	Unit Management Observation	(66.7 percent)

In summary, the experimental group exceeded the acceptable level of 85 percent agreement on seven of the eight sources, whereas the control group achieved or exceeded the acceptable level on only four of the eight sources. Neither group achieved the acceptable level of percentage agreement on the Patient Record source.

### Significant Differences and Hypotheses Testing

All observations combined. As shown in Table 4, a 2 x 2 table Chi-square analysis, .05 level of significance, was performed on the total frequency of disagreements to the total frequency of agreements for both the experimental and the control group. The Chi-square value was significant  $\chi^2 (1, N = 3308) = 12.308, P < .05$ .

In addition, a t-test of differences between the means of the experimental group's observation scores ( $M = 87.2$ ) and the control group's observation scores ( $M = 82.8$ ) was performed. The t-test, at .05 level of significance (two tailed), indicated that the experimental group (nonspecialized observers) had a significantly higher interrater reliability score than did the control group (specialized observers):  $t (df 28) = 2.24, P < .05$ .

Therefore, the null hypothesis which stated, "There are no differences in the interrater reliability scores between specialized and nonspecialized observers," may be rejected on the basis of the significant difference between the means of the experimental and control groups.

By objective. As shown in Table 4, a Chi-square analysis, .05 level of significance, of the experimental and control groups' frequency of disagreements to agreements on criteria related to each objective was performed. The only objective which had a significant Chi-square value was Objective 2 (Physical Needs Are Attended),  $\chi^2 (1, N = 771) = 13.819, P < .05$ . For Objective 2, the experimental group had 32 disagreements out of 387 criteria, whereas the control group had 66 disagreements out of 384 criteria.

**TABLE 4**  
**DIFFERENCES BETWEEN GROUPS BY OBJECTIVES**  
**ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Objective</u>	<u>a <math>\chi^2</math> (df=1)</u>
(1) Nursing Care Plan is Formulated	0.151
(2) Physical Needs Are Attended	13.819 *
(3) Nonphysical Needs Are Attended	1.555
(4) Nursing Care Objectives Are Evaluated	0.000
(5) Unit Procedures Followed	1.443
(6) Delivery of Nursing Care is Facilitated	1.891
All Objectives	12.308 *

**Note:** Level of Significance .05.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $p < .05$ .

In summary, a significant difference was found between the experimental and control groups' frequency of disagreements related to Objective 2. Therefore, the null hypothesis which stated, "There are no differences in the interrater reliability scores between specialized and nonspecialized observers," was again rejected on the

basis of the significant difference between the groups' disagreements related to Objective 2.

**By source of information.** As shown in Table 5, a Chi-square analyses, .05 level of significance, of the experimental and control groups' disagreements to agreements on criteria collected by each of the eight different sources was performed. Of the eight sources, two had significant Chi-square values: Source V (Nursing Observation),  $\chi^2$  (1,  $N = 192$ ) = 5.832,  $P = <.05$ , and Source VI (Environmental Observation),  $\chi^2$  (1,  $N = 403$ ) = 7.142,  $P = <.05$ ). For Source V, the experimental group had 7 disagreements out of 97 criteria, whereas the control group had 18 disagreements out of 95 criteria. For Source VI, the experimental group had 13 disagreements out of 192 criteria, whereas the control group had 32 disagreements out of 211 criteria.

In summary, significant differences were found between the experimental and control groups' frequency of disagreements related to Source V (Nursing Observation) and Source VI (Environmental Observation). Therefore, the null hypothesis which stated, "There are no differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference, and record abstraction methods," may be rejected on the basis of the significant differences between the groups' disagreements related to the Nursing Observation and Environmental Observation sources.

**Sources within each group.** As mentioned previously, neither the experimental nor the control group achieved the minimum acceptable

TABLE 5.  
DIFFERENCES BETWEEN GROUPS BY SOURCE  
ESTIMATED BY CHI-SQUARE ANALYSES

<u>Sources</u>	<u>a <math>\chi^2</math> (df=1)</u>
I Patient Record	1.703
II Patient Observation	3.175
III Patient Interview	0.089
IV Nursing Interview	1.501
V Nursing Observation	5.832 *
VI Environmental Observation	7.142 *
VII Observer Inference	0.102
VIII Unit Management Observation	2.700
All Sources Combined	12.308 *

Note: Yates correction used for all expected cells less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values:  $P < .05$ .

level of 85 percent agreement for data collected from the Patient Record source. Therefore, for both groups a Chi-square analyses, .05 level of significance was performed on only the disagreements and agreements associated with the Patient Record source to the disagreements and agreements associated with each of the other seven sources.

For the experimental group, as shown in Table 6, significant Chi-square values were obtained when the Patient Record source was compared with each of five other sources:

Source II, Patient Observation ( $\chi^2$ , (1,  $N$  = 622) = 15.150,  $P$  = <.05)  
 Source III, Patient Interview ( $\chi^2$ , (1,  $N$  = 882) = 21.838,  $P$  = <.05)  
 Source IV, Nursing Interview ( $\chi^2$ , (1,  $N$  = 886) = 40.093,  $P$  = <.05)  
 Source V, Nursing Observation ( $\chi^2$ , (1,  $N$  = 627) = 12.758,  $P$  = <.05)  
 Source VI, Environmental Observation ( $\chi^2$ , (1,  $N$  = 722) = 24.906,  $P$  = <.05)

Nonsignificant Chi-square values were obtained for the experimental group when the Patient Record source was compared with Source VII (Observer Inference),  $\chi^2$  (1,  $N$  = 550) = 2.690,  $P$  >.05 and with Source VIII (Unit Management Observation),  $\chi^2$  (1,  $N$  = 542) = 2.401,  $P$  >.05 .

For the control group, as shown in Table 7, significant Chi-square values were obtained when the Patient Record source was compared with each of four other sources:

Source II, Patient Observation ( $\chi^2$ , (1,  $N$  = 618) = 4.798,  $P$  = <.05)  
 Source III, Patient Interview ( $\chi^2$ , (1,  $N$  = 874) = 28.438,  $P$  = <.05)  
 Source IV, Nursing Interview ( $\chi^2$ , (1,  $N$  = 911) = 40.502,  $P$  = <.05)

**TABLE 6**  
**DIFFERENCES WITHIN EXPERIMENTAL GROUP'S SOURCES**  
**ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Source I</u>	<u>vs.</u>	<u>Other Sources</u>	<u>a <math>\chi^2</math></u>
I Patient Record f = 123		II Patient Observation f = 5	15.150 *
I Patient Record f = 123		III Patient Interview f = 38	21.838 *
I Patient Record f = 123		IV Nursing Interview f = 25	40.093 *
I Patient Record f = 123		V Nursing Observation f = 7	12.758 *
I Patient Record f = 123		VI Environmental Observation f = 13	24.906 *
I Patient Record f = 123		VII Observer Inference f = 1	2.690
I Patient Record f = 123		VIII Unit Observation f = 0	2.401

**Note:** Yates correction used when expected cells less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $P < .05$



TABLE 7

DIFFERENCES WITHIN CONTROL GROUP'S SOURCES  
ESTIMATED BY CHI-SQUARE ANALYSES

<u>Source I</u>	<u>vs.</u>	<u>Other Sources</u>	<u>a <math>\chi^2</math></u>
I Patient Record f = 145		II Patient Observation f = 11	4.798 *
I Patient Record f = 145		III Patient Interview f = 33	28.438 *
I Patient Record f = 145		IV Nursing Interview f = 35	40.502 *
I Patient Record f = 145		V Nursing Observation f = 18	2.528
I Patient Record f = 145		VI Environmental Observation f = 11	11.180 *
I Patient Record f = 145		VII Observer Inference f = 2	3.124
I Patient Record f = 145		VIII Unit Observation f = 4	0.035

Note: Yates correction used when expected cells less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $P < .05$ .

Source VI, Environmental Observation ( $\chi^2$ , (1,  $N$  = 755) = 11.180,  $P$  = <.05)

Nonsignificant Chi-square values were obtained for the control group when the Patient Record source was compared with Source V, (Nursing Observation)  $\chi^2$ , (1,  $N$  = 639) = 2.528,  $P$  = >.05, Source VII, (Observer Inference)  $\chi^2$ , (1,  $N$  = 568) = 3.124,  $P$  = >.05, and Source VIII, (Unit Management Observation)  $\chi^2$ , (1,  $N$  = 556) = .035,  $P$  = >.05.

In summary, for both groups, the frequency of disagreements associated with the Patient Record source were significantly different than the frequency of disagreements associated with the Patient Observation, Patient Interview, Nursing Interview, and Environmental Observation sources. As well, for the experimental group, the Patient Record source was also significantly different than the Nursing Observation source. Therefore, the null hypothesis which stated, "There are no differences in interrater reliability scores for data collected by the interview, observation, inference, or record abstraction methods within either the specialized or nonspecialized observer groups," may be rejected on the basis of the significant differences within each group's disagreements related to various sources.

### Reasons for Disagreements

As discussed in Chapter III, each pair's comments related to questions in disagreement were subsequently coded into one of nine categories representing reasons for disagreement. The following is an analyses of these reasons.

TABLE 8  
EXPERIMENTAL AND CONTROL GROUPS' REASONS  
ESTIMATED BY PERCENTAGE

<u>Reason</u>	<u><sup>a</sup>Experimental Group (Percent)</u>	<u><sup>b</sup>Control Group (Percent)</u>
	N = 212	N = 285
1. Error in Recording	15.6	17.2
2. Patient Response Unclear	6.6	3.9
3. Nurse Response Unclear	3.3	2.8
4. Question Unclear	5.2	3.5
5. Standard Varies	2.4	2.5
6. Judgement Varies	22.6	33.3
7. Careless Reading of Question	20.3	15.8
8. Patient Record Confusing	17.0	17.2
9. Other Reasons	7.1	3.9
Total Responses	1651	1657

**Note:** Rounding rule: 0-4 down, 5-9 up.

a = experimental group composed of 15 pairs of nonspecialized observers.

b = control group composed of 15 pairs of specialized observers.

Reasons as a percent of total disagreement. As shown in Table 8, the experimental group, based on 15 paired observations, had a total of 212 disagreements out of 1651 responses. The control group, based on 15 paired observations, had a total of 285 disagreements out of 1657 responses.

For the experimental group, 75.5 percent of their total disagreements were a result of the following four reasons:

Judgement Varies	(22.6 percent)
Careless Reading of Question	(20.3 percent)
Patient Record Confusing	(17.0 percent)
Error in Recording	(15.6 percent)

For the control group, 83.5 percent of their total disagreements were a result of the following four reasons:

Judgement Varies	(33.3 percent)
Patient Record Confusing	(17.2 percent)
Error in Recording	(17.2 percent)
Careless Reading of Question	(15.8 percent)

Differences between groups' reasons. As shown in Table 9, a Chi-square analysis was conducted which compared the experimental group with the control group on the frequency of disagreements associated with each of the nine reasons. A significant Chi-square value was obtained for reason six (Judgement Varies),  $\chi^2 (1, N = 3308) = 15.968, P < .05$ . The experimental group had a frequency of 48 disagreements, out of 1651 total responses, attributed to Judgement Varies, whereas the control group had a frequency of 95 disagreements, out of 1657 total responses, attributed to Judgement Varies.

TABLE 9  
DIFFERENCES BETWEEN GROUPS' REASONS  
ESTIMATED BY CHI-SQUARE ANALYSES

<u>Reasons</u>	<u><sup>a</sup>Experimental Group (f)</u>	<u><sup>b</sup>Control Group (f)</u>	<u><sup>c</sup> <math>\chi^2</math></u>
1. Error in Recording	33	49	3.142
2. Patient Response Unclear	14	11	0.374
3. Nurse Response Unclear	7	8	0.032
4. Question Unclear	11	10	0.052
5. Standard Varies	5	7	0.080
6. Judgement Varies	48	95	15.968 *
7. Careless Reading of Question	43	45	0.040
8. Record Confusing	36	49	1.993
9. Other Reasons	15	11	0.63
Total Responses	1651	1657	

**Note:** Yates correction used when expected cell less than 10.

a = experimental group composed of 15 pairs of nonspecialized observers.

b = control group composed of 15 pairs of specialized observers.

c =  $\chi^2$  (df=1) critical value = >3.841.

\* = significant  $\chi^2$  values.  $P < .05$ .

Differences within groups' reasons. As noted previously in Table 8, Judgement Varies was the reason which accounted for the highest percentage of disagreements for both the experimental and control group (22.6 percent and 33.3 percent respectively). Therefore, a Chi-square analyses was conducted for each group, which compared only the frequency of disagreements associated with Judgement Varies to the frequency of disagreements associated with each of the other eight reasons.

For the experimental group, as shown in Table 10, significant Chi-square values were obtained when Judgement Varies was compared with five other reasons:

Patient Response Unclear  $(\chi^2, (1, N = 424) = 21.839 \underline{P} = <.05)$

Nurse Response Unclear  $(\chi^2, (1, N = 424) = 35.119 \underline{P} = <.05)$

Question Unclear  $(\chi^2, (1, N = 424) = 26.954 \underline{P} = <.05)$

Standard Varies  $(\chi^2, (1, N = 424) = 39.871 \underline{P} = <.05)$

Other Reasons  $(\chi^2, (1, N = 424) = 20.302 \underline{P} = <.05)$

Nonsignificant Chi-square values were obtained when Judgement Varies was compared with three other reasons:

Error in Recording  $(\chi^2, (1, N = 424) = 3.434 \underline{P} = >.05)$

Careless Reading  $(\chi^2, (1, N = 424) = 0.350 \underline{P} = >.05)$

Record Confusing  $(\chi^2, (1, N = 424) = 2.138 \underline{P} = >.05)$

In summary, within the experimental group's reasons, the frequency of disagreements associated with Judgement Varies did not appear to be significantly different than the frequency of disagreements associated with Error in Recording, Careless Reading, and Patient Record Confusing.

**TABLE 10**  
**DIFFERENCES WITHIN EXPERIMENTAL GROUP'S REASONS**  
**ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Reason 6</u>	<u>vs.</u>	<u>Other Reasons</u>	<u>a <math>\chi^2</math></u>
Judgement Varies f = 48		Error in Recording f = 33	3.434
Judgement Varies f = 48		Patient Response unclear f = 14	21.839 *
Judgement Varies f = 48		Nurse Response unclear f = 7	35.119 *
Judgement Varies f = 48		Question Unclear f = 11	26.954 *
Judgement Varies f = 48		Standard Varies f = 5	39.871 *
Judgement Varies f = 48		Careless Reading f = 43	0.350
Judgement Varies f = 48		Record Confusing f = 36	2.138
Judgement Varies f = 48		Other Reasons f = 15	20.302 *

**Note:** Yates correction used when expected cell less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $p < .05$ .

For the control group, as shown in Table 11, significant Chi-square values were obtained when Judgement Varies was compared with each of the other eight reasons:

Error in Recording	$(\chi^2, (1, N = 570) = 19.662 \underline{P} = <.05)$
Patient Response Unclear	$(\chi^2, (1, N = 570) = 81.773 \underline{P} = <.05)$
Nurse Response Unclear	$(\chi^2, (1, N = 570) = 92.128 \underline{P} = <.05)$
Question Unclear	$(\chi^2, (1, N = 570) = 84.347 \underline{P} = <.05)$
Standard Varies	$(\chi^2, (1, N = 570) = 92.469 \underline{P} = <.05)$
Careless Reading	$(\chi^2, (1, N = 570) = 23.671 \underline{P} = <.05)$
Record Confusing	$(\chi^2, (1, N = 570) = 19.662 \underline{P} = <.05)$
Other Reasons	$(\chi^2, (1, N = 570) = 81.773 \underline{P} = <.05)$

In summary, within the control group's reasons, the frequency of disagreements associated with Judgement Varies appeared to be significantly different than the frequency of disagreements associated with each of the other reasons.

### **Objective 2 (Physical Needs Are Attended)**

As noted previously, Table 4, when the experimental and control groups' disagreements to agreements for each objective were estimated by Chi-square analyses, the only objective that had a significant Chi-square value was Objective 2 (Physical Needs Are Attended). The following is an analyses of this objective by subobjective, source, and reasons.

**Objective 2 by subobjective.** A Chi-square analyses was conducted to compare the experimental with the control groups' disagreements to total agreements for each subobjective relative to Objective 2. As shown in Table 12, a significant Chi-square value was obtained



TABLE 11

**DIFFERENCES WITHIN CONTROL GROUP'S REASONS  
ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Reason 6</u>	<u>vs.</u>	<u>Other Reasons</u>	<u>a <math>\chi^2</math></u>
Judgement Varies f = 95		Error in Recording f = 49	19.662 *
Judgement Varies f = 95		Patient Response f = 11	81.773 *
Judgement Varies f = 95		Nurse Response f = 8	92.128 *
Judgement Varies f = 95		Question Unclear f = 10	84.347 *
Judgement Varies f = 95		Standard Varies f = 7	92.469 *
Judgement Varies f = 95		Careless Reading f = 45	23.671 *
Judgement Varies f = 95		Record Confusing f = 49	19.662 *
Judgement Varies f = 95		Other Reasons f = 11	81.773 *

**Note:** Yates correction used when expected cell less than 10.

a =  $\chi^2$ (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $P < .05$ .

**TABLE 12**  
**DIFFERENCES BETWEEN GROUPS BY SUBOBJECTIVE (OBJECTIVE 2)**  
**ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Subobjective</u>	<u>a <math>\chi^2</math></u>
2.1 (Patient Protected from Accident and Injury)	3.224
2.2 (Need For Physical Rest and Comfort Attended)	0.097
2.3 (Need For Physical Hygiene Attended)	3.805
2.4 (Need For Supply of O <sub>2</sub> Attended)	1.292
2.5 (Need For Activity Attended)	0.285
2.6 (Nutrition and Fluid Balance Attended)	0.656
2.7 (Elimination Attended)	0.186
2.8 (Need For Skin Care Attended)	0.087
2.9 (Patient Protected From Infection)	4.208 *

**Note:** Yates correction used when expected cell less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values. P <.05.

for Subobjective 9 (Patient Protected from Infection),  $\chi^2 = (1, N = 117) = 4.208, P = <.05$ . The experimental group had 2 disagreements out of 61 questions associated with Subobjective 9, whereas the control group had 9 disagreements out of 56 questions associated with Subobjective 9.

In summary, between the experimental and control groups' disagreements related to the various subobjectives of Objective 2, the control groups' disagreements associated with Subobjective 9 appeared to be significantly different than the experimental groups' disagreements. No significant differences were found between the groups' disagreements related to the other subobjectives.

Objective 2 by source of information. As noted previously, Table 5, when the experimental and control groups' disagreements per source of information were compared, based upon 15 observations and all objectives combined, significant Chi-square values were obtained for Source V (Nursing Observation) and Source VI (Environmental Observation). Therefore, specifically in relation to Objective 2, a Chi-square analysis was conducted to compare the experimental group with the control groups' frequency of disagreements associated with each source of information.

As shown in Table 13, a significant Chi-square value was obtained for Source VI (Environmental Observation),  $\chi^2 (1, N = 237) = 11.243, P = <.05$ . For Objective 2, the experimental group had 4 out of 32 disagreements associated with Source VI, whereas the control group had 21 out of 66 disagreements associated with Source VI.

In summary, between the experimental and control groups'

TABLE 13

**DIFFERENCES BETWEEN GROUPS' BY SOURCES (OBJECTIVE 2)  
ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Sources</u>	<u>Experimental Group (f)</u>	<u>Control Group (f)</u>	<u><sup>a</sup> <math>\chi^2</math></u>
1. Patient Record	10	18	3.470
2. Patient Observation	5	11	3.175
3. Patient Interview	9	6	0.162
4. Nursing Interview	0	4	1.658
5. Nursing Observation	3	6	0.464
6. Environmental Observation	4	21	11.243*
7. Observer Inference	1	0	0.003
8. Unit Management Observation	NA	NA	
Total Disagrees	32	66	

**Note:** Yates correction used when expected cells less than 10.

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $P < .05$ .

**TABLE 14**  
**DIFFERENCES BETWEEN GROUPS' BY REASONS (OBJECTIVE 2)**  
**ESTIMATED BY CHI-SQUARE ANALYSES**

<u>Reasons</u>	<u>Experimental Group (f)</u>	<u>Control Group (f)</u>	<u>a <math>\chi^2</math></u>
1. Error in Recording	7	14	2.455 <sup>a</sup>
2. Patient Response Unclear	4	2	0.160
3. Nurses Response Unclear	0	0	0.0
4. Question Unclear	2	5	0.592
5. Standard Varies	0	0	0.0
6. Judgement Varies	7	24	9.852 *
7. Careless Reading of Question	7	11	0.536
8. Patient Record Confusing	2	8	2.572
9. Other Reasons	3	2	0.0
Total Reasons	32	66	
Objective Response	387	384	

**Note:** Yates correction used when expected cells less than 10

a =  $\chi^2$  (df=1) critical value = >3.841.

\* = Significant  $\chi^2$  values.  $P < .05$ .

disagreements related to the various sources for Objective 2, the control group's disagreements associated with the Environmental Observation source appeared to be significantly different than the experimental group's disagreements.

Reasons within Objective 2. As shown in Table 14, a Chi-square analyses conducted to compare the experimental with the control groups' frequency of disagreements associated with each reason relative to Objective 2, a significant Chi-square value was obtained for reason six (Judgement Varies),  $\chi^2 (1, N = 771) = 9.852, P = >.05$ . The experimental group had 7 out of 32 disagreements, associated with Objective 2, attributed to Judgement Varies, whereas the control group had 24 out of 66 disagreements, associated with Objective 2, attributed to Judgement Varies.

In summary between the experimental and control groups' various reasons for disagreement associated with Objective 2, the control group's disagreements associated with Judgement Varies appeared to be significantly different than the experimental group's disagreements.

## CHAPTER V

### SUMMARY, DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

#### SUMMARY

In Canadian health care organizations, there is widespread use of data collecting instruments which require the use of raters, observers, or interviewers. In Canadian hospitals, this is a result of growing concern about, and attention to, quality assurance activities. This growing concern comes not only from the recent guidelines of the Canadian Council on Hospital Accreditation, which stipulate that each hospital must monitor the quality of services provided, but also from increasing pressure to manage resources in ways which are both morally and legally defensible. Accordingly, it is of utmost importance to achieve acceptable levels of reliability with these instruments in order to ensure consistent, accurate data. However, minimal research has been done on the reliability or the factors affecting the reliability of these quality monitoring instruments.

The social science literature notes that there are a multitude of factors that might affect the reliability of a quality monitoring instrument. For instruments which require the use of raters (observers, interviewers), factors relating to the instrument, the rater, and the respondent need to be considered in order to achieve acceptable levels of reliability. For instruments of this nature, interrater reliability, an estimate of the degree to which two or more independent raters are consistent in their judgements, is recommended as the most critical reliability criterion. The reliability of an instrument

is not a fixed property of the instrument, but instead it is a condition of the instrument that needs to be established in each setting where the instrument is being used.

In nursing departments of Canadian acute care hospitals, there is an increasingly widespread use of ongoing data collecting instruments which require the use of raters in an attempt to provide objective measures of the quality of nursing care. However, the current variations in practice relating to factors that might affect the reliability of these instruments, combined with the current minimal research in this area, raises many questions regarding the utility of these measures.

A survey, conducted in conjunction with this study, of twenty Canadian nursing departments utilizing the Rush-Medicus Quality Monitoring Instrument, found numerous variations in practice relative to factors that might affect the reliability of the instrument. Specifically related to raters, differences were found regarding the numbers of trained observers, the observer's position in the organization, the assignment plan for observers, the training and retraining process for observers, and the interrater reliability testing practice.

In this study, the relationship of two specific factors, observer specialization and data collection methods, to interrater reliability when using the Rush-Medicus Quality Monitoring Instrument was examined. Observer specialization, an independent variable, was defined as the degree nurse observers conduct observations (audits) in their areas of current clinical experience and consisted of two elements:



specialized observers (those who conduct audits only in their area of current clinical experience) and nonspecialized observers (those who conduct audits outside their area of current clinical experience). Data collection methods, another independent variable, was defined as the existing sources of information specified for each criterion in the Rush-Medicus Instrument and consisted of eight elements: patient record abstraction, patient interview, nurse interview, patient observation, nursing observation, environmental observation, unit management observation and observer inference. Interrater reliability, the dependent variable, was defined as the degree to which two or more independent raters were consistent in their judgements and consisted of one measure: percentage agreement scores. The following null hypotheses were tested:

- 1 Ho: There are no differences in the interrater reliability scores between specialized and nonspecialized observers.
- 2 Ho: There are differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference, and record abstraction methods.
- 3 Ho: There are no differences in interrater reliability scores for data collected by the interview, observation, inference, and record abstraction methods within either the specialized or the nonspecialized observers.

This study, conducted in a large tertiary-care teaching hospital, was a field experiment utilizing the Posttest - Only Control Group design. The Rush-Medicus Quality Monitoring Instrument, currently

utilized and validated at the study hospital, was used. This instrument consists of 440 questions relating to 357 criteria. The criteria are grouped into homogenous clusters to assess six objectives and 28 subobjectives related to the nursing process. The six nursing objectives are:

1. Nursing care plan is formulated.
2. Physical needs are attended
3. Nonphysical needs are attended.
4. Nursing care objectives are evaluated.
5. Unit procedures are followed.
6. Delivery of nursing care is facilitated.

A representative sample of each subobjectives criteria has been randomly assigned to a series of questionnaires relevant to patient type and to clinical service. For the purpose of this study, each questionnaire was considered equivalent and therefore was not held constant between patients.

The study population consisted of the existing population of 170 specialized nurse observers at the study hospital. Each of the nurse observers at the study hospital is assigned to conduct audits (observations) only within the clinical area in which he/she has current clinical experience. For example, medical nurse observers audit only on medical nursing units. This population of specialized observers was stratified into seven subgroups representing the existing clinical service areas. A simple random selection of 62 observers (subjects) was drawn from within the various subgroups. Each of the subjects selected consented to participate in the study. The subjects selected

from within each subgroup were randomly assigned to pairs. The resulting 31 pairs were then randomly assigned to either experimental or control group status. This random assignment process yielded 15 experimental pairs and 16 control pairs.

Over a three month period (January - March, 1984), each of the experimental pairs was assigned to conduct one interrater reliability observation, using the Rush-Medicus Instrument, on a randomly selected nursing unit outside their existing clinical subgroup and thereby outside their area of current auditing and clinical nursing experience. This assignment of subjects to a randomly selected nursing unit outside their existing subgroup constituted the experimental treatment for the experimental pairs.

Over the same three months, each of the control pairs was assigned to conduct one interrater reliability observation, using the Rush-Medicus Instrument, on a randomly selected nursing unit within their existing clinical subgroup and thereby within their area of current auditing and clinical nursing experience.

Each interrater reliability observation (audit) performed by either an experimental or control pair consisted of the two observers randomly selecting two patients on their assigned unit and independently, but simultaneously, conducting the required patient and unit observations utilizing the appropriate Rush-Medicus questionnaires. For the control pairs, each audit required approximately 1½ - 2 hours to complete and, for the experimental pairs, each audit took approximately 2 - 2½ hours to complete.

Immediately following the completion of the audit, each pair

attended a debriefing session conducted by the researcher. This debriefing session included a calculation of the pair's percentage agreement score based on all three questionnaires plus a discussion of each of the questions found to be in disagreement. The debriefing sessions<sup>o</sup> lasted approximately 1 hour per pair.

The study extended into the month of April to accommodate a few pairs of observers. One control group pair was eventually excluded because of the subjects inability to find any mutually convenient time to audit. One experimental pair was required to repeat their interrater reliability observation on another randomly assigned unit because of an instance of discussion related to the patient record. This example of contamination necessitated the introduction of additional measures to ensure that independent, yet simultaneous, observations were conducted. A total of 15 observations, comprised of 30 patients and 15 unit observations, were completed by each group.

A special computer program was designed to accommodate the analyses required by this study. For both the experimental and control group, a percentage of agreement statistic was calculated for all the nursing process objectives combined, for each objective individually, and for criteria related to each source of information specified in the Rush-Medicus Instrument. The null hypotheses were tested using the t-test and the Chi-square statistic. In addition, each pair's comments regarding questions in disagreement were coded into one of nine categories representing reasons for disagreement. This coding frame was not subjected to formal reliability testing. An analysis of the reasons for disagreement between and within the group's was performed

using percent and the Chi-square statistic. The Yates correction was used whenever appropriate.

The findings of this study were as follows:

Percentage agreement. The percentage of agreement statistic, calculated for all objectives combined, for each objective, and for each source of information, was compared to the 85 percent minimum standard recommended for the interrater reliability of the Rush-Medicus Quality Monitoring Instrument. For all the objectives combined, this comparison found that the experimental group's (nonspecialized observers) mean score of 87.2 percent interrater reliability, based on 15 paired observations, exceeded the minimum standard. However, of these 15 paired observations, four nonspecialized pairs achieved less than 85 percent interrater reliability on their individual audits. For the control group (specialized observers) this comparison found that the specialized observers' mean score of 82.8 percent interrater reliability, also based on 15 observations, did not meet the minimum standard. As well, of these 15 paired observations, nine specialized pairs achieved less than 85 percent interrater reliability on their individual audits (Table 1).

In regard to the interrater reliability scores for each of the six objectives, this comparison found that neither the nonspecialized observers nor the specialized observers achieved the minimum standard for two objectives: Nursing Care Plan is Formulated and Nursing Care Objectives are Evaluated. In addition, the specialized observers did not achieve the minimum standard for two other objectives: Physical Needs are Attended and Unit Procedures Followed (Table 2).

With regard to the interrater reliability scores for criteria related to each of the eight sources of information, this comparison found that neither the nonspecialized observers nor the specialized observers achieved the minimum standard for criteria obtained from the Patient Record source. In addition, the specialized observers did not achieve the minimum standard for criteria obtained from three other sources: Nursing Observation, Environmental Observation, and Unit Management Observation (Table 3).

**Hypotheses testing.** The first null hypothesis which stated, "There are no differences in the interrater reliability scores between specialized and nonspecialized observers," may be rejected on the basis of a significant t-test and a significant Chi-square value (Table 4). These findings suggest that the nonspecialized observers had significantly higher interrater reliability scores than did the specialized observers. Furthermore, Chi-square analyses, comparing the nonspecialized and specialized groups' disagreements to agreements on each objective, found a significant Chi-square value related to Objective 2, Physical Needs are Attended, (Table 4). This finding suggests that the nonspecialized observers had significantly fewer disagreements associated with Objective 2 than did the specialized observers.

Specifically related to objective 2, a Chi-square analyses comparing each group's disagreements to total agreements for each subobjective comprising Objective 2, found a significant Chi-square value associated with Subobjective 9, Patient Protected from Infection, (Table 12). This finding suggests that the nonspecialized observers

had significantly fewer disagreements associated with Subobjective 9 than did the specialized observers.

The second null hypothesis which stated, "There are no differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference, and record abstraction methods," may be rejected on the basis of significant Chi-square values found between the nonspecialized and specialized observers' disagreements associated with two sources: Nursing Observation and Environmental Nursing Observation (Table 5). These findings suggest that the nonspecialized observers had significantly fewer disagreements associated with these two sources than did the specialized observers.

Specifically related to Objective 2, the one objective previously noted to be significantly different between the two groups, a Chi-square analyses comparing the nonspecialized observers' disagreements to the specialized observers' disagreements for each source of information related to Objective 2, found a significant Chi-square value associated with the Environmental Observation source (Table 13). This finding suggests that the nonspecialized observers had significantly fewer disagreements associated with the Environmental Observation source of information than did the specialized observers.

The third null hypothesis which stated, "There are no differences in interrater reliability scores for data collected by the interview, observation, inference, or record abstraction methods within either the specialized or nonspecialized observer groups," may be rejected on the basis of significant Chi-square values obtained when each group's

disagreements related to the Patient Record source was compared to their disagreements related to the remaining seven sources. For the nonspecialized observers, significant Chi-square values were obtained when the Patient Record source was compared to the following five sources (Table 6):

1. Patient Observation
2. Patient Interview
3. Nursing Interview
4. Nursing Observation
5. Environmental Observation

These findings suggest that the nonspecialized observers' frequency of disagreements associated with the Patient Record source was significantly greater than their frequency of disagreements associated with five other sources. Conversely, these findings suggest that nonspecialized observers' disagreements associated with the Patient Record source were similar (i.e., nonsignificant Chi-square values) to their disagreements associated with the remaining two sources: Observer Inference and the Unit Management Observation.

For the specialized observers, significant Chi-square values were obtained when the Patient Record source was compared to the following four sources (Table 7):

1. Patient Observation
2. Patient Interview
3. Nursing Interview
4. Environmental Observation

These findings suggest that the specialized observers' frequency



of disagreements associated with the Patient Record source was significantly greater than their frequency of disagreements associated with four other sources. Conversely, these findings suggest that the specialized observers' disagreements associated with the Patient Record source were similar (i.e., nonsignificant Chi-square values) to their disagreements associated with the three remaining sources: Observer Inference, Unit Management Observation, and Nursing Observation.

**Reasons for disagreements.** Of the nine reasons for disagreements categorized in this study, Judgement Varies, Careless Reading of the Question, Patient Record Confusing, and Error in Recording were the four major reasons for each group's disagreements. For the nonspecialized observers, these four reasons accounted for 75.5 percent of their total 212 disagreements. For the specialized observers, these four reasons accounted for 83.5 percent of their total 285 disagreements (Table 8).

A Chi-square analysis, comparing each group's frequency of disagreement attributed to each of the nine reasons, found a significant Chi-square value associated with the Judgement Varies reason (Table 9). This finding suggests that the nonspecialized observers had significantly fewer disagreements attributed to Judgement Varies than did the specialized observers. Conversely, for each of the remaining eight reasons, the nonspecialized and specialized observers had similar frequencies of disagreement (i.e., nonsignificant Chi-square values).

Within each group, a Chi-square analysis comparing the frequency of disagreement attributed to Judgement Varies was compared to the

frequency of disagreement attributed to each of the other eight reasons. For the nonspecialized observers, significant Chi-square values were found when Judgement Varies was compared with five other reasons (Table 10):

1. Patient Response Unclear
2. Nurse Response Unclear
3. Question Unclear
4. Standard Varies
5. Other Reasons

This finding suggests that, for the nonspecialized observers, significantly more disagreements were attributed to Judgement Varies than to Patient Response Unclear, Nurse Response Unclear, Question Unclear, Standard Varies, and Other Reasons. Conversely, the frequency of disagreements attributed to Judgement Varies was similar to the frequency of disagreements attributed to the remaining three reasons: Error in Recording, Careless Reading of the Question, and Patient Record Confusing (i.e., nonsignificant Chi-square values).

For the specialized observers, significant Chi-square values were found when Judgement Varies was compared with eight other reasons (Table 11):

1. Error in Recording
2. Patient Response Unclear
3. Nurse Response Unclear
4. Question Unclear
5. Standard Varies
5. Careless Reading

7. Record Confusing

8. Other Reasons

This finding suggests, that for the specialized observers, significantly more disagreements were attributed to Judgement Varies than to any other reason.

Specifically related to Objective two (Physical Needs are Attended), a Chi-square analyses comparing each group's reasons for disagreements associated with Objective two, found that Judgement Varies was again significantly different (Table 14). This finding suggests that for Objective two (the one objective found to be significantly different between the two group's) the nonspecialized observers had significantly fewer disagreements attributed to Judgement Varies than did the specialized observers.

In summary, these findings suggest that Judgement Varies was the major reason for the specialized observers significantly greater number of disagreements and thus their significantly lower interrater reliability score.

All of the findings in this field experiment were and are subject to a number of limitations and delimitations. Generalizability is limited for the following reasons:

1. The study was conducted in only one hospital.
2. The hospital is a tertiary-care teaching hospital and thereby represents a small portion of Canadian hospitals.
3. The study examined only one quality monitoring instrument.
4. The researcher's definition of specialized and nonspecialized observers.

5. The hospital's current method of assigning observers.
6. The subject's consent to participate may have produced "Hawthorne-like" effects.
7. The experimental treatment may have produced "reactive arrangements" effects on the nonspecialized observers.
8. The researcher's categorization of reasons for disagreements.

Internal validity may have been jeopardized by the following reasons:

1. Random selection was expected to provide equivalence between groups, however there is a slight chance that randomization failed and the groups would differ even if no treatment intervened.
2. The effect of not holding constant the clinical services, nursing units, patients, questionnaires, and observer pairs is unknown.
3. The potential contamination effects during independent observations may not have been controlled by the measures taken in the study.
4. The lack of formal reliability testing on the coding frame for reasons for disagreements may limit internal validity.

## DISCUSSION

This discussion on the study results is presented in the following order: similar findings between the groups and significant differences between the groups are discussed and compared with other research findings, limitations of the study are discussed, and implications for practice are presented.

### Similar Findings Between the Groups

A minimum standard of 85 percent agreement is recommended for the interrater reliability of the Rush-Medicus Quality Monitoring Instrument. This standard has generally been adopted by users of this instrument.

All objectives combined. In regard to all the objectives combined, based on 15 observations, which included 30 patients and 15 unit assessments per group, the nonspecialized observers' mean score (87.2 percent) exceeded this standard, whereas the specialized observers' mean score (82.8 percent) did not achieve this standard.

However, the mean score, for either group, is considerably lower than the average of 90 percent agreement reported by ten of the twenty survey hospitals on the results of their regular, ongoing interrater reliability testing. Similarly, for the specialized observers, their mean score (82.8 percent) was considerably lower than the study hospital's mean score of 87.4 percent agreement found on interrater reliability testing associated with new observers.

In regard to interrater reliability estimates on individual audits, based on all the objectives combined, the nonspecialized observers' percentage agreement scores ranged between 77 and 98 percent, which

included four audits that did not achieve the minimum standard. The specialized observers' percentage agreement scores ranged between 77 and 95 percent, which included nine audits that did not achieve the minimum standard. The above ranges are both wider than the range of 83 - 92 percent reported by Hegyvary and Hausmann (1976) on the results of their initial interrater reliability testing of the Rush-Medicus Quality Monitoring Instrument.

The above individual audits for each group which did not achieve the recommended standard of 85 percent agreement (i.e., 4/15 and 9/15), are similar to the findings of Ventura et al. (1980) using the I.C.C. standard of .75 and the Rush-Medicus Instrument. They reported that one pair of raters did not achieve the standard in 3 of 5 instances, while the other pair did not achieve the standard in 2 of 5 instances during the first period of testing.

By objective. In regard to interrater reliability estimates for each objective in the Rush-Medicus Instrument, neither the nonspecialized nor that specialized observers achieved the minimum standard of 85 percent on criteria related to two objectives: Objective 1, Nursing Care Plan is Formulated, (77.6 percent and 76.2 percent respectively) and Objective 4, Achievement of Objectives in Evaluated (65.6 percent and 65.6 percent respectively). A comparison of this finding with other research findings using percent agreement per objective and the Rush-Medicus Instrument is not available. However, Ventura et al. (1980), using the I.C.C. statistic and the Rush-Medicus Instrument, reported that neither of the two pairs of raters achieved the recommended I.C.C. standard of .75 related to

Objective 1 during the first period of testing, and only one pair exceeded the standard in the second period of testing. In relation to Objective 4, they reported that one pair achieved the I.C.C. standard during the first period of testing, whereas neither pair achieved the standard during the second period of testing.

By source. In relation to Objectives 1 and 4, it is important to note that the criteria associated with these objectives almost exclusively require the rater to obtain information from the Patient Record source (e.g., 83 of the 85 criteria related to Objective 1, and all of the 22 criteria related to Objective 4). Accordingly, the above finding is consistent with the study finding that neither the nonspecialized nor the specialized observers achieved the minimum standard of 85 percent agreement for criteria obtained from the Patient Record source (76.8 percent and 73.4 percent respectively). Furthermore, this finding, which suggests both groups of observers had difficulty with criteria obtained from the Patient Record source, is consistent with the finding of Herman and Cayten (1980) who found low interrater reliability on medical record abstraction was associated with variables which require the rater to use judgement. In the researcher's opinion, many of the Rush-Medicus criteria, obtained from the Patient Record source, require the observer (rater) to use judgement.

In addition to the Patient Record source not achieving the minimum standard of 85 percent interrater reliability for both groups of observers, findings from the Chi-square analyses, which compared each group's frequency of disagreements associated with criteria related

to the Patient Record source with the frequency of disagreements associated with criteria related to each of the other sources, suggest similar proportions of disagreement were experienced by both groups with criteria related to the Observer Inference and the Unit Management Observation sources of information (i.e., nonsignificant Chi-square values). However, these findings are inconsistent with the percentage agreement scores obtained by both groups on these sources. For example, both groups exceeded the minimum standard for criteria related to the Observer Inference source (95.0 percent and 91.7 percent respectively). Related to the Unit Management Observation source, the nonspecialized observers achieved 100 percent whereas the specialized observers achieved only 66.7 percent. Therefore, these findings, which suggest similar difficulties were experienced with the Observer Inference, the Unit Management Observation, and the Patient Record sources must be viewed with caution. The apparent inconsistency may be attributed to the limitation of the Chi-square analysis when dealing with disproportionate cell entries.

By reason. In regard to the reasons for disagreements, Judgement Varies, Careless Reading of the Question, Patient Record Confusing, and Error in Recording accounted for the majority of the disagreements experienced by both the nonspecialized and specialized observers (75.5 percent and 85.5 percent respectively). It is interesting to note that of these reasons, Judgement Varies accounted for the largest percentage of disagreement for both groups (22.6 percent and 33.3 percent for the nonspecialized and specialized observers respectively). This finding is consistent with the previous finding



that criteria obtained from the Patient Record source pose difficulty for both groups. It is also consistent with the finding of Herman and Cayten (1980) who found low interrater reliability on medical record abstraction was associated with variables which require the rater to use judgement.

It is also of interest to note that for both groups between 33 and 36 percent of their total disagreements were related to what may be termed observer carelessness. For example, for the nonspecialized observers, disagreements related to Careless Reading of the Question and Error in Recording accounted respectively for 20.3 percent and 15.6 percent of their total disagreements. Similarly, for the specialized observers, Careless Reading of the Question and Error in Recording accounted respectively for 15.8 percent and 17.2 percent of their total disagreements. These findings raise numerous questions; particularly questions about observer selection, length of tenure, retraining, frequency of interrater reliability testing, and conditions complicating the observer's task. In this regard, Steinkamp (1966) reported that an effective interviewer scored higher on the Edwards Personal Performance Schedule in areas of dominance, self-confidence, and attention to detail. However, Hausmann, Hegyvary and Newman (1976) reported that the personal characteristics of observers, as measured by the California Psychological Inventory (CPI) and the Watson - Glaser Inventory, do not provide an adequate basis for selection. Instead, they concluded that extensive training sessions and reliability testing at least every month were the significant factors in the reliable use of the Rush-Medicus Instrument.

In summary, these findings suggest that nurse observers at the study hospital:

1. achieved lower interrater reliability scores than reported by the survey hospitals;
2. achieved a wider range of percentage agreement scores than reported at the time of the initial testing of the instrument;
3. failed to achieve the minimum standard for interrater reliability on 13 out of 30 observations;
4. failed to achieve the minimum standard for Objective 1 and 4;
5. failed to achieve the minimum standard for criteria related to the Patient Record source of information; and,
6. attributed the majority of their disagreements to Judgement Varies, Careless Reading of the Question, Patient Record Confusing, and Error to Recording.

#### Significant Differences Between the Groups

Ho 1. The first hypothesis which stated, "There are no differences in the interrater reliability scores between specialized and nonspecialized observers," was rejected on the basis of the significant values (t-test and Chi-square) obtained when comparisons of the two group's scores and total disagreements were calculated. These findings suggest that the nonspecialized observers had significantly better interrater reliability scores and significantly fewer total disagreements than did the specialized observers. Furthermore, a comparison of disagreements per objective found a significant Chi-square value associated with Objective 2. This finding suggests that among the six objectives, the nonspecialized observers had significantly

fewer disagreements associated with Objective 2 (Patients Physical Needs Are Attended).

These findings, which suggest the nonspecialized observer had significantly higher interrater reliability scores, supports the practice of 80 percent of the hospitals surveyed who reported assigning nurses to conduct audits on any nursing unit (except their own). It is also consistent with the previous finding that the nonspecialized observers' mean score exceeded the minimum standard whereas the specialized observers' mean score did not. However, it is inconsistent with the argument that, due to the high degree of technical and clinical complexity in contemporary nursing practice, nurse observers conducting observations in clinical areas on which they have current clinical experience ought to have a higher probability of consistent measurement. Since no previous research has been reported in this area, a comparison with other findings is not possible.

Ho 2. The second null hypothesis which stated, "There are no differences between the interrater reliability scores of specialized and nonspecialized observers for data collected by the interview, observation, inference, and record abstraction methods," was rejected on the basis of significant Chi-square values obtained between the two groups related to two sources of information: Nursing Observation and Environmental Observation. This finding suggests that the nonspecialized observers had significantly better interrater reliability associated with these two sources than did the specialized observers.

This finding is consistent with the percentage agreement scores obtained by each group on these two sources (Nursing

Observation - 92.8 percent and 81.1 percent respectively, and Environmental Observation - 93.2 percent and 84.8 percent respectively). Accordingly, this finding again supports the practice of those hospitals who assign observers to all nursing units. Once again, however, it is inconsistent with the argument that observers with current clinical experience in the area ought to have a higher probability of consistent measurement. Since no previous research has been reported in this area a comparison with other findings is not possible.

Ho 3. The third hypothesis which stated, "There are no differences in interrater reliability scores for data collected by the interview, observation, observer inference, or record abstraction methods within either the specialized or the nonspecialized observer groups," was rejected on the basis of significant Chi-square values obtained when each group's frequency of disagreements associated with the Patient Record source was compared to their frequency of disagreement on each of the other sources. For the nonspecialized observers, their frequency of disagreements associated with the Patient Record source was significantly different (i.e., greater) than their frequency of disagreement associated with five other sources. For the specialized observers, their frequency of disagreements associated with the Patient Record source was significantly different (i.e., greater) than their frequency of disagreement related to four other sources.

Conversely, these findings suggest that, for the nonspecialized observers, the Patient Record source was similar, in terms of disagreements experienced, to the Observer Inference source and the Unit Management Observation source. Similarly, for the specialized

observers, these findings suggest that the Patient Record source was similar, in terms of disagreements experienced, to Observer Inference, Unit Management Observation, and Nursing Observation sources of information.

The above findings are, however, inconsistent with the percentage agreement scores related to the sources. For example, for the nonspecialized observers, only the criteria related to the Patient Record source did not meet the minimum standard. For the specialized observers, criteria related to the Patient Record source, Unit Management Observation source, Nursing Observation source, and Environmental Observation source did not meet the minimum standard. Therefore, these findings must be viewed with caution. The apparent inconsistency may be attributed to the limitation of the Chi-square analysis when dealing with disproportionate cell entries.

Significant differences among reasons. A comparison of each group's frequency of disagreement associated with each reason, found a significant Chi-square value associated with the Judgement Varies reason. Furthermore, a comparison of each group's reasons for disagreement associated with Objective 2 (Patient's Physical Needs are Attended) again found a significant value associated with Judgement Varies. These findings suggest that the nonspecialized observers had significantly fewer disagreements related to Judgement Varies than did the specialized observers.

In summary, these findings suggest that the nonspecialized observers (those who conduct audits outside their area of current clinical experience) had:

1. significantly fewer disagreements;
2. significantly higher interrater reliability scores;
3. significantly fewer disagreements associated with the Nursing Observation and Environmental Observation sources of information; and,
4. significantly fewer disagreements related to the Judgement Varies reason.

Conversely, these findings suggest that the specialized observers (those who conduct audits within their area of current clinical experience) had:

1. significantly more disagreements;
2. significantly lower interrater reliability scores;
3. significantly more disagreements associated with the Nursing Observation and Environmental Observation sources of information; and,
4. significantly more disagreements related to the Judgement Varies reason.

Whereas these findings support the practice of the majority of the hospital's surveyed who assign nurse observers to audit throughout the hospital, they are inconsistent with the argument (logic) that observers with current clinical experience in their area ought to have a higher probability of consistent measurement. One can only speculate as to possible explanations for these unexpected findings.

On the one hand, in relation to specialized observers (those who conduct audits within their area of current clinical experience) a possibility exists that because of their familiarity and current

clinical expertise in the area, they respond with differing, individualized standards to the criteria. This possibility may be supported by the significant differences noted between the groups on Objective 2 (Physical Needs are Attended), Nursing Observation and Environmental Observation sources of information, and disagreements attributed to Judgement Varies.

Another possibility may be that the specialized observers were less motivated, compared to the nonspecialized observers, to perform the interrater reliability audit. Although they too consented to be subjects, the actual task required was not out of their normal routine. This possibility may be supported by the researcher's observation that few, if any, specialized observers responded to their assignment with apprehension.

Still another possibility may be that specialized observers conducting observations on units which they are familiar present less objectivity and more bias.

On the other hand, in relation to the nonspecialized observers (those who conduct audits outside their area of current clinical experience), a possibility exists the randomization failed, and the groups were different before the experimental treatment occurred. This possibility is not supported by the comparison on selected variables between the two groups following random selection (p. 30).

Another possibility may be that nonspecialized observers conducting observations on units to which they are unfamiliar are more objective.

Still another possibility may be that a "reactive arrangements" effect was produced by the experimental treatment. Campbell and

Stanley (1966, p. 21) note that "reactive arrangements" are induced by the subject's knowledge that he is participating in an experiment. Furthermore, he suggests that similar effects could be induced by the presentation of an experimental treatment if it is an out - of - the ordinary event. In this study, whereas both groups consented to participate, it was only the nonspecialized group who were required to perform an out - of - the ordinary event, that is, the interrater reliability audit on an unfamiliar nursing unit outside their area of current clinical experience. This possibility may be supported by the researcher's observation that many nonspecialized observers reacted to their assignment to an unfamiliar area with apprehension and caution. Comments such as, "It wasn't as bad as I thought it would be," and, "We didn't want to let you down," were frequently stated by nonspecialized observers during their debriefing sessions. The possibility of a "reactive arrangements" effect may be further supported by the fact that, on average, the nonspecialized observers took longer to complete their audit. The strong possibility that a "reactive arrangements" effect occurred limits the generalizability of these findings. To avoid this effect, Campbell and Stanley (1966) suggest that, in much research on teaching methods, "... there is no need for the students to know that an experiment is going on" (p. 21). However, in nursing, this is usually not feasible. It is especially not feasible in light of the Canadian Nurses Association and the study hospital's guidelines on research which stress the consent of the subject to participate.



### Limitations

In addition to the possibility that a "reactive arrangements" effect occurred among the nonspecialized observers, other possible threats to external validity further limit generalization of the study results to other settings. These include the following:

1. The study was conducted in only one hospital.
2. The hospital is a tertiary - care teaching hospital and thereby represents a small portion of Canadian hospitals.
3. The study examined only one quality monitoring instrument.
4. The hospital's specific methods for selection, training, retraining, and assigning nurse observers may have influenced the findings.
5. The subjects consent to participate may have produced Hawthorne-like effects.
6. The researcher's definition of specialized and nonspecialized observers may have influenced the findings.
7. The researcher's categorization of reasons for disagreements may have influenced the findings.

Internal validity may have been jeopardized because of the following reasons:

1. The random selection of subjects was expected to provide equivalence between groups, however there is a slight chance that randomization failed and the groups would differ even if no treatment intervened.
2. The clinical services, nursing units, patients, questionnaires, and pairs of observers were not held constant, instead they were

allowed to vary to approximate the normal, quality monitoring process in nursing departments.

3. The measures taken in this study may not have prevented contamination during independent observations:
4. The lack of formal interrater reliability testing of the coding frame may have influenced the categorization of reasons.

### Implications

Because of the possible threats to external validity associated with this study, the implications flowing from the study's results are directed solely towards the study hospital.

Specialized versus nonspecialized observers. The nonspecialized observers' significantly higher interrater reliability scores, and their significantly fewer disagreements associated with the Nursing Observation source, the Environmental Observation source, and the Judgement Varies reason, may imply, dependent on the explanation chosen, that nurse observers ought to be assigned to audit on any nursing unit. Therefore, the study hospital's current method of assigning nurse observers only to nursing units within their area of current clinical experience needs to be carefully examined. Because of the alternate explanations proposed for this unexpected finding, further research on this variable is needed.

Interrater reliability by objective. The finding that neither the specialized nor the nonspecialized observer groups achieved the minimum standard of 85 percent agreement for Objectives 1 and 4 implies serious problems with the criteria related to these objectives. Therefore, the actual scores nursing units receive relative to these objectives

must be viewed with caution until the problems are identified and corrected.

In addition, the specialized observers did not achieve the minimum standard for two other objectives and for all the objectives combined. This finding suggests that the study hospitals interrater reliability associated with the quality monitoring audit is below the recommended standard. Therefore, the reliability of the audit data must be viewed with caution until the interrater reliability estimates are improved.

Interrater reliability by source. The finding that neither group achieved the minimum standard of 85 percent agreement for criteria related to the Patient Record source implies serious problems associated with obtaining data from the patient record. Accordingly, this finding assists in identifying the nature of the problem associated with Objectives 1 and 4 because virtually all of the criteria related to these two objectives requires information be obtained from the patient record. Therefore, the study hospital needs to concentrate on improving the interrater reliability of criteria obtained from the patient record. To accomplish this both the criteria and the conditions related to the Patient Record source need to be examined and improved wherever possible. For example, each of the 105 criterion related to the Patient Record source needs to be reviewed and revised for clarity, as well as for the addition of appropriate standards to facilitate reliability. This review can be assisted by the study's findings which identified reasons for disagreement relative to each criterion (Appendix L). The condition of the patient record itself needs to be examined to ensure it facilitates the abstraction of information. Factors

surrounding the condition of the patient record are many and diverse. They range from convenience of the nursing station, to the design of forms to accommodate information entry and retrieval, to the clarity of data entry, to the type of record holder. Regarding the latter, the study hospital's patient record is kept vertical fashion on a clipboard. In the past, both medical and nursing staff have identified numerous difficulties in the recording and abstraction of information from this type of record holder. Accordingly, it is conceivable that the nurse observers experienced similar difficulties obtaining information from the patient record, and moreover, that these difficulties are reflected in the low interrater reliability scores associated with the Patient Record source.

In addition to the Patient Record source, the findings suggest that the specialized observers also experienced difficulty with the Unit Management Observation source of information. However, this finding is less significant because of the fact that only one criterion requires information be obtained from the Unit Management Observation source as opposed to the 105 criteria which require information be obtained from the Patient Record source. Nevertheless, an examination of the criterion for clarity as well as for the addition of an appropriate standard is warranted.

Reasons for disagreement. The finding that the majority of the reasons for disagreement were attributed to Judgment Varies, Careless Reading of the Question, Patient Record confusing, and Error in Recording suggests that interrater reliability estimates, and thus confidence in the quality monitoring measurements, can be improved by concentrating

on these four reasons. Therefore, an examination and study of the numerous factors that may contribute to these reasons is essential. At the study hospital, factors such as observer selection, length of observer appointments, training and retraining practices, frequency of interrater reliability testing, control measures taken during interrater reliability, and criteria and conditions related to the Patient Record source of information are among the numerous factors that need to be reviewed, possibly modified, and certainly evaluated.

## CONCLUSIONS AND RECOMMENDATIONS

This study examined the relationship of observer specialization and of data collection methods to interrater reliability when using a quality monitoring instrument. The findings, due to the limitations of the study, cannot be generalized to other settings. Instead, the results are specific to the study hospital and to the Rush-Medicus Instrument.

### Observer Specialization and Interrater Reliability

The findings suggest that a relationship between observer specialization and interrater reliability may exist. The unexpected finding that nonspecialized observers achieved significantly higher interrater reliability estimates than specialized observers is subject to a number of alternate explanations. Of these possible explanations, not the least is the strong possibility that a "reactive arrangements" effect occurred within the experimental group. Further research on this variable is needed.

### Recommendations for further research include:

- A study comparing the interrater reliabilities of nurse observers who are typically assigned to any nursing unit with the interrater reliabilities of nurse observers who are typically assigned only to areas of current clinical experience.
- A study on the interrater reliability of the Rush-Medicus Instrument that is designed to reduce the potential for Hawthorne effects and "reactive arrangements" effects.
- A similar study in a hospital which typically assigns nurse observers to any nursing unit within the hospital.

- A study on the interrater reliability of the Rush-Medicus Instrument that is designed to further reduce the potential for contamination during interrater reliability testing.

#### **Data Collection Methods and Interrater Reliability**

The findings suggest that a relationship between different data collection methods and interrater reliability may exist. The finding that neither group achieved the minimum standard of 85 percent agreement for criteria obtained from the Patient Record source suggests this source presents the most difficulty to nurse observers using this quality monitoring instrument. Furthermore, the finding that nurse observers did not achieve the minimum standard for criteria related to Objectives 1 and 4 is consistent with the above concern about the Patient Record source, since virtually all the criteria related to these two objectives require information be obtained from the patient record. Accordingly, efforts directed towards improving the criteria and/or conditions related to the Patient Record source could increase interrater reliability estimates and thus increase confidence in the quality measurements obtained.

#### **Recommendations for further research include:**

- Studies on the interrater reliability of the Rush-Medicus Instrument following specific efforts directed towards improving the criteria and/or conditions related to the Patient Record source of information.

#### **Reasons for Disagreements**

The study found that the majority of disagreements experienced by nurse observers were related to four reasons: Judgement Varies,

Patient Record Confusing, Careless Reading of the Question, and Error in Recording. Of these reasons, Judgement Varies accounted for 22-23 percent of the total disagreements and, what may be termed, observer carelessness accounted for another 33-36 percent.

In regard to quality monitoring instruments used on an ongoing basis, these findings raise serious questions about observer selection, observer motivation, observer tenure, and observer retraining. Steinkamp (1966) found that effective interviewers scored higher on the Edwards Personal Reference Schedule in areas of dominance, self-confidence, and attention to detail. However, Hausmann, Hegyvary, and Newman (1976) concluded that extensive training and monthly interrater reliability testing were more significant for the reliability of the Rush-Medicus Instrument than was observer selection. Research on all these factors is needed. It is especially needed because in many institutions, which employ large numbers of nurse observers, the recommendation that monthly interrater reliability testing of nurse observers be conducted is frequently unfeasible, impractical, and too costly. Research is needed on alternate methods, such as videos and simulated laboratory settings, for use in the retraining and interrater reliability testing of nurse observers involved with ongoing quality monitoring instruments.

**Recommendations for further research include:**

- A study examining the relationship of observer characteristics and interrater reliability of the Rush-Medicus Quality Monitoring Instrument.



• A study on the interrater reliability of nurse observers following retraining sessions utilizing videos and laboratory settings.

• A study comparing interrater reliability testing of nurse observers in the real setting with interrater reliability testing of nurse observers in a simulated laboratory setting.

## REFERENCES

- Bartko, J.J., and Carpenter, W.T. (1976). On the methods and theory of reliability. Journal of Nervous and Mental Disease, 163(5), 307-317.
- Boyd, H.T., and Westfall, R. (1970). Interviewer bias once more revisited. Journal of Marketing Research, 7, 249-253.
- Broyles, R.W., and Lay, C.M. (1979). Statistics in health administration (Vol. 1). Germantown: Aspen Systems Corporation.
- Campbell, D.T., and Stanley, J.C. (1963). Experimental and quasi-experimental designs in research. Boston: Houghton Mifflin Company.
- Canadian Council on Hospital Accreditation. (1977). Guide to hospital accreditation.
- Canadian Council on Hospital Accreditation. (1983). Standards for accreditation of canadian health care facilities.
- Ferguson, G.A. (1966). Statistical analysis in psychology and education. New York: McGraw-Hill.
- Fox, D.J. (1972). Fundamentals of research in nursing. Norwalk: Appleton-Century Crofts.
- Giovannetti, P. (1981). Aspects of measurement. In Y.M. Williamson (Ed.), Research methodology and its application to nursing, pp. 145-164. New York: John Wiley and Sons.
- Goodwin, L.D., and Prescott, P.A. (1981). Issues and approaches to estimating interrater reliability in nursing research. Research in Nursing and Health, 4, 323-337.
- Hausmann, D., Hegyvary, S., and Newman, J. (1976). Monitoring quality of nursing care: Part II Assessment and study of correlates. (DHEW Publication No. (HRA) 76-7). Bethesda, Maryland: U.S. Department of Health, Education, and Welfare.
- Hegyvary, S.T., and Hausmann, R.K. (1976). Monitoring nursing care quality. Journal of Nursing Administration, 6, 3-9.
- Herman, N., and Cayten, C. (1980). Interobserver and intraobserver reliability in the collection of emergency medical services data. Health Services Research, 15(2), 127-143.
- Horn, B.J. (1980). Establishing valid and reliable criteria. Nursing Research, 29(2), 88-93.

- Jelinek, R.C., Hausmann, R.K., Hegyvary, S.T., and Newman, J.F. (1975). A methodology for monitoring quality of nursing care (DHEW Publication No. (HRA) 74-25). Bethesda, Maryland: U.S. Department of Health, Education, and Welfare.
- Kahn, R.L., and Cannell, C.F. (1963). The dynamics of interviewing. New York: John Wiley and Sons, Inc.
- Kidder, L.H. (1981). Research methods in social relations. New York: Holt, Rinehart, and Winston, Inc.
- Kerlinger, F.N. (1973). Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart, and Winston, Inc.
- Maguire, T.O., and Hazlett, C.B. (1969). Reliability for the researcher. Alberta Journal of Educational Research, 15, 117-125.
- Medicus Systems Corporation (undated). Orientation manual for nursing quality monitoring methodology.
- Mickevicius, V., and Stoughton, W.V. (1984). Management and quality assurance. Health Management Forum, 5(3), 4-13.
- Polit, D.F., and Hungler, B.P. (1978) Nursing research: Principles and methods. New York: J.B. Lippincott Company.
- Steinkamp, S.W. (1966). Some characteristics of effective interviewers. Journal of Applied Psychology, 50(6), 487-492.
- Stuart, A. (1968). Basic ideas of scientific sampling. London: Charles Griffin and Company Limited.
- Sudman, S., and Bradburn, N. (1974). Response effects in surveys. Chicago: Aldine Publishing Company.
- Ventura, M.T., Hageman, P.T., Slaker, M.J., and Fox, R.M. (1980). Interrater reliabilities for two measures of nursing care quality. Research in Nursing and Health, 3, 25-32.
- Ventura, M.T., and Hageman, P.T. (1978). Testing for the reliability, validity, and sensitivity of quality of nursing care measures: Final report. Washington, D.C.: Health Services Research and Development Service, Veterans Administration.

**APPENDIX A**

**DEMOGRAPHIC CHARACTERISTICS OF SURVEY HOSPITALS**

## APPENDIX A

## DEMOGRAPHIC CHARACTERISTICS OF SURVEY HOSPITALS

(20 Canadian Hospitals using the Rush-Medicus Quality Monitoring Instrument)

<u>Number of Beds:</u>	<u>Percentage</u>
100 - 199	05%
200 - 299	05%
300 - 499	25%
500 or more	65%

Services Provided:

General:

Medical and Surgical	95%
Obstetrics	70%
Intensive Care	95%
Coronary Care	95%
Pediatrics	55%
Urological	90%
Gynecological	95%
Neurological	65%

Special:

Pediatrics	40%
Convalescent	20%
Rehabilitative	55%
Chronic	40%
Isolation	35%
Geriatrics	40%

Services Provided:

Special:

Orthopedic	80%
Alcoholic	30%
Arthritic	45%
Mental Retardation	10%
Psychiatry	90%
Other	35%

**APPENDIX B**

**SURVEY QUESTIONNAIRE**



**QUESTIONNAIRE TO QUALITY ASSURANCE COORDINATORS ON  
INTER-OBSERVER RELIABILITY AND THE  
RUSH-MEDICUS QUALITY MONITORING INSTRUMENT**

Please do not  
write in this  
column cc

- |     |  |      |
|-----|--|------|
| 1.  | What is the bed size (beds set up for use, excluding bassinets for newborns) of your hospital? (Please check one of the following) | 1, 2 |
| 1.  | 50 - 99 ( )  |      |
| 2.  | 100 - 199 ( )  |      |
| 3.  | 200 - 299 ( )  | 3    |
| 4.  | 300 - 499 ( )  |      |
| 5.  | 500 - + ( )  |      |
| 2.  | Which of the following services does your hospital provide? (Please check appropriate services)                                    |      |
|     | General  |      |
| 1.  | Medical and Surgical ( )   | 4    |
| 2.  | Obstetrical ( )  | 5    |
| 3.  | Intensive Care ( )   | 6    |
| 4.  | Coronary Care ( )  | 7    |
| 5.  | Pediatrics ( )   | 8    |
| 6.  | Urological ( )   | 9    |
| 7.  | Gyneocological ( )   | 10   |
| 8.  | Neurosurgical ( )  | 11   |
|     | Special  |      |
| 9.  | Pediatrics ( )   | 12   |
| 10. | Convalescent ( )   | 13   |
| 11. | Rehabilitation ( )   | 14   |
| 12. | Chronic ( )  | 15   |
| 13. | Extended Care ( )  | 16   |
| 14. | Isolation ( )  | 17   |
| 15. | Geriatrics ( )   | 18   |
| 16. | Orthopedic ( )   | 19   |
| 17. | Alcoholic ( )  | 20   |
| 18. | Arthritic ( )  | 21   |
| 19. | Mental Retardation ( )   | 22   |
| 20. | Psychiatric ( )  | 23   |
| 21. | Other ( )  | 24   |



3. How long have you used the Rush-Medicus Methodology for Monitoring Quality of Nursing Care? (Please check one of the following)

- 1. less than 12 months ( )
- 2. 13 - 24 months ( )
- 3. 25 - 26 months ( )
- 4. 37 - 48 months ( )
- 5. 49 + - ( )

25

4. How many audit tours do you schedule, on average, each month? (Please check one of the following)

- 1. 1 - 20 ( )
- 2. 21 - 40 ( )
- 3. 41 - 60 ( )
- 4. 61 - 80 ( )
- 5. 81 - 100 ( )
- 6. 101 - 120 ( )
- 7. 121 - 140 ( )
- 8. Other ( )

26

If other, please explain \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

5. What percentage of scheduled tours are you able to complete, on average, each month? (Please check appropriate percentage)

- 1. 1 - 10% ( )
- 2. 11 - 20% ( )
- 3. 21 - 30% ( )
- 4. 31 - 40% ( )
- 5. 41 - 50% ( )
- 6. 51 - 60% ( )
- 7. 61 - 70% ( )
- 8. 71 - 80% ( )
- 9. 81 - 90% ( )
- 10. 91 - 100% ( )

27, 28

6. How many trained nurse-observers do you currently have?

- 1. 1 - 10 ( )
- 2. 11 - 20 ( )
- 3. 21 - 30 ( )
- 4. 31 - 40 ( )
- 5. 41 - 50 ( )
- 6. 51 - 60 ( )
- 7. 61 - 70 ( )
- 8. 71 - 80 ( )
- 9. 81 - 90 ( )
- 10. 91 + - ( )

29, 30

7. How many of your trained nurse-observers represent the following classifications? (Please place numbers in space provided)

- 1. General Duty Nurse (graduate) ( )
- 2. General Duty Nurse (registered) ( )
- 3. Assistant Charge Nurse ( )
- 4. Charge Nurse ( )

31, 32  
33, 34  
35, 36  
37, 38

- 5. Clinical Instructor ( )
- 6. Supervisor ( )
- 7. Others ( )

39, 40  
41, 42  
43, 44

If others, please explain \_\_\_\_\_

---



---



---

8. Are your nurse-observers expected to conduct audits on all nursing units involved in the audit program (except their own)?

- 1. Yes ( )
- 2. No ( )

45

If no, please explain \_\_\_\_\_

---



---



---

9. How are nurse-observers selected in your hospital? (Please check appropriate answers)

- 1. Volunteer ( )
- 2. Required of their position ( )
- 3. Full-time nurse-observer positions ( )

46  
47  
48

9a. If nurse-observers volunteer to audit in your hospital, do you have a selection criteria?

- 1. Yes ( )
- 2. No ( )

49

If yes, briefly describe or attach a copy of the selection criteria please \_\_\_\_\_

---



---



---

9b. If you have full-time nurse-observer positions, what are your selection criteria? (Please briefly explain or attach job description) \_\_\_\_\_

---



---



---



---

10. What is your training process for nurse-observer? (Please check one of the following processes)

1. Classroom review of the Medicus orientation manual and practice session in class. ( )

2. Classroom review of the Medicus orientation manual and practice session in class, followed by an inter-observer reliability tour. ( )

50

3. Classroom review of the Medicus orientation manual and short practice session, plus a practice session with an experienced observer in the real setting, followed by an inter-observer reliability tour. ( )

4. Other ( )

If other, please describe \_\_\_\_\_

---



---



---



---

11. Do you assign an experienced nurse-observer along with the trainee nurse-observer on the trainees' first inter-observer tour?

- 1. Yes ( )
- 2. No ( )

51

12. What is the average length (in hours) of your training process for the nurse-observers? (Include class time, practice sessions, and inter-observer tour)

- 1. less than 1 hour ( )
- 2. 1 - 4 hours ( )
- 3. 5 - 8 hours ( )
- 4. 9 - 16 hours ( )
- 5. 16 + hours ( )

52

13. Do you conduct inter-observer reliability tours for each nurse-observer subsequent to the tour associated with the orientation training process?

- 1. Yes ( )
- 2. No ( )

53

If yes, please complete 13a, 13b, and 13c.

13a. How frequently do you conduct inter-observer reliability tours for each nurse-observer subsequent to the tour associated with the orientation training process?

- 1. quarterly ( )
- 2. semiannually ( )
- 3. annually ( )
- 4. other ( )

54

If other, please explain \_\_\_\_\_

---



---



---

13b. Do you assign one "expert", experienced nurse-observer to each inter-observer tour subsequent to the orientation training process?

- 1. Yes ( )
- 2. No ( )

55

13c. In the past 12 months, what has been your average inter-observer reliability score for all inter-observer tours?

( )

56, 57

14. Do you conduct training sessions for experienced nurse-observers?

- 1. Yes ( )
- 2. No ( )

58

If yes, please describe the retraining process and the frequency of this process. \_\_\_\_\_

---

---

---

---

---

15. Do you have a committee that deals with Quality Monitoring scores and processes? (eg. Nursing Audit Committee)

- 1. Yes ( )
- 2. No ( )

59

If yes, please attach terms of reference and membership.

Thank you for your cooperation.

**APPENDIX C**  
**AUDIT SCHEDULE**

APPENDIX C  
AUDIT SCHEDULE

January / April / July / October	15 Stations	110 Audits
February / May / August / November	18 Stations	114 Audits
March / June / September / December	19 Stations	127 Audits

<u>Area</u>	<u>Station</u>	<u>Tours</u>
-------------	----------------	--------------

January / April / July / October

<u>OBS./GYNE</u>	60	10
	62	6
	63	10
	64	10
	65	10
	69	6
	70	6

<u>BERHART</u>	81	6
	82	6
	83	6
	84	6

SPECIAL AREAS

EMERGENCY	20	6
BURN UNIT	49	6
PARR	50	10
ICU	68	6

<u>Area</u>	<u>Station</u>	<u>Tours</u>
<u>February / May / August / November</u>		
<u>MEDICINE</u>	5A2 (02)	6
	5A4 (03)	6
	44	6
	52	6
<u>MEDICINE</u>	40	6
	41	6
	66	6
<u>MEDICINE</u>	31	6
	32	6
	43	6
<u>PEDS</u>	33	8
	35	6
	36	6
	37	10
	38	6
<u>VETS HOME</u>	11	6
	12	6
	13	6



<u>Area</u>	<u>Station</u>	<u>Tours</u>
<u>March / June / September / December</u>		
<u>PSYCH</u>	14	6
	23	6
	24	6
<u>SURGERY</u>	4C3 (21)	6
	4C4/D2 (22)	6
	46	6
	47	9
	54	6
	56	6
	57	9
	58	6
<u>SURGERY</u>	5C2 (04)	6
	5C3 (05)	6
	5C4/D2 (06)	6
	4A2 (08)	6
	34	6
	42	6
	51	10
55	9	

**APPENDIX D**

**RUSH-MEDICUS QUALITY MONITORING INSTRUMENT  
NURSING CARE OBJECTIVES**

**APPENDIX D**  
**NURSING CARE OBJECTIVES**

<u>Score</u>	<u>Unit</u>
1.1	Condition is Assessed on Admission
1.2	Data Relevant to Care are Ascertained
1.3	Current Condition is Assessed
1.4	Written Care Plan is Formulated
1.5	Nursing Plan is Co-ordinated with Medical Plan
1.0	NURSING CARE PLAN IS FORMULATED
2.1	Patient is Protected from Accident and Injury
2.2	Need for Comfort and Rest is Attended
2.3	Need for Physical Hygiene is Attended
2.4	Need for Supply of Oxygen is Attended
2.5	Need for Activity is Attended
2.6	Need for Nutrition and Fluid Balance is Attended
2.7	Need for Elimination is Attended
2.8	Need for Skin Care is Attended
2.9	Patient is Protected from Infection
2.0	PATIENT'S PHYSICAL NEEDS ARE ATTENDED
3.1	Patient is Oriented to Hospital Facilities on Admission
3.2	Patient is Extended Courtesy by Staff
3.3	Patient's Privacy and Civil Rights are Honoured
3.4	Psycho-Emotional Well-Being is Attended (Interpersonally)
3.5	Patient is Taught Health Maintenance/Illness Prevention
3.6	Patient's Family is Included in the Care Process
3.7	Psycho-Emotional Well-Being is Attended (Therapeutically)
3.0	PATIENT'S NON-PHYSICAL NEEDS ARE ATTENDED
4.1	Records Document Care Provided
4.2	Patient's Response to Therapy is Evaluated
4.0	ACHIEVEMENT OF OBJECTIVES IS EVALUATED
5.1	Isolation and Decontamination Procedures are Followed

- 5.2 Unit is Prepared for Emergency Situations
- 5.3 Medical-Legal Procedures are Followed
- 5.4 Safety and Protective Procedures are Followed,
- 5.0 UNIT PROCEDURES ARE FOLLOWED FOR PATIENT PROTECTION
- 6.1 Nursing Report Follows Prescribed Standards
- 6.2 Nursing Management is Provided
- 6.3 Clerical Services are Provided
- 6.4 Environment and Housekeeping Services are Provided
- 6.5 Professional and Administrative Services are Provided
- 6.0 DELIVERY OF NURSING CARE IS FACILITATED

Blank Score - Indicates No Valid Responses  
Asterisk (\*) - Indicates Insufficient Valid Responses for a  
Reliable Score

**APPENDIX E**

**RUSH-MEDICUS QUALITY MONITORING QUESTIONNAIRES**

## APPENDIX E

## RUSH-MEDICUS QUALITY MONITORING QUESTIONNAIRES

## Exhibit B: Questionnaire Number by Series by Clinical Area

Clinical Area	Patient Type				Unit
	1	2	3	4	
Emergency Department	111	121			151
	112	122			152
	113	124			153
Labor and Delivery	211	221	231		251
	212	222	232		252
	213	223	233		253
	214	224	234		
	215		235		
			236		
Psychiatry	311				351
	312				352
	313				353
	314				
	315				
	316				
Nursery	421	431	441	451	
	422	432	442	452	
	423	433	443	453	
	424	434	444		
	425	435	445		
	426	436	446		
Parents		427	437	447	
		428	438	448	
		429	439	449	
General Care	511	521	531	541	551
	512	522	532	542	552
	513	523	533	543	553
	514	524	534	544	
	515	525	535	545	
		526	536	546	
		527	537	547	
			538		
	611			651	

**APPENDIX F**

**RUSH-MEDICUS QUALITY MONITORING INSTRUMENT  
NUMBER OF QUESTIONS BY SOURCE OF INFORMATION**

## APPENDIX F

RUSH-MEDICUS QUALITY MONITORING INSTRUMENT  
NUMBER OF QUESTIONS BY SOURCE OF INFORMATION

TOTAL 440 QUESTIONS

Source #1	Patient Record	195 Questions
Source #2	Patient Observation	19 Questions
Source #3	Patient Interview	82 Questions
Source #4	Nursing Personnel Interview	73 Questions
Source #5	Nursing Personnel Observation	21 Questions
Source #6	Patient Environment Observation	45 Questions
Source #7	Observer Inference	4 Questions
Source #8	Unit Management Observation	1 Question



**APPENDIX G**  
**STUDY SCHEDULE**

**APPENDIX G**  
**STUDY SCHEDULE**

January

OBS/GYN

Experimental Pair	Station	64
Control Pair	Station	65
Control Pair	Station	64

Long Term Care

Experimental Pair	Station	84
Experimental Pair	Station	82
Control Pair	Station	81
Control Pair	Station	81

Special Areas

Experimental Pair	Station	20
Experimental Pair	Station	49
Control Pair	Station	68
Control Pair	Station	20

FebruaryMedicine

Experimental Pair	Station	43
Experimental Pair	Station	66
Control Pair	Station	43
Control Pair	Station	40
Control Pair	Station	5A2

Pediatrics

Experimental Pair	Station	36
Experimental Pair	Station	33
Experimental Pair	Station	36
Experimental Pair	Station	37
Control Pair	Station	35
Control Pair	Station	33

MarchPsychiatry

Experimental Pair	Station	14
Experimental Pair	Station	23
Control Pair	Station	23
Control Pair	Station	23

Surgery

Experimental Pair	Station	404
Experimental Pair	Station	42
Control Pair	Station	42
Control Pair	Station	55
Control Pair	Station	57

**APPENDIX H**  
**CONSENT FORM**

APPENDIX H  
VARIABLES ASSOCIATED WITH INTERRATER RELIABILITY OF A  
QUALITY MONITORING INSTRUMENT

CONSENT FORM

I, \_\_\_\_\_ agree to participate in a research study on interrater reliability conducted by Dorothy Meilicke.

It is my understanding that:

- 1) my participation in the study is voluntary and I will be able to withdraw from the study at any time without penalty;
- 2) my name will not be made known in any publication or communication;
- 3) my participation will involve one interrater reliability tour; and,
- 4) I may be assigned to perform an interrater reliability tour on any nursing unit at the University of Alberta Hospitals.

Signature: \_\_\_\_\_

Witness: \_\_\_\_\_

Date: \_\_\_\_\_

**APPENDIX I**  
**TESTING FOR INTER-OBSERVER RELIABILITY**



## APPENDIX I

## TESTING FOR INTER-OBSERVER RELIABILITY

Inter-observer reliability is the level of agreement among observers. When used in Research, reliability refers to the repeatability of observations; that is, what percentage of the time do two observers collecting data from the same source at the same time agree on what they observe. Inter-observer reliability is an extremely important issue of any data collection. If the observers consistently do not agree, the data may not be useful.

The following are steps to be completed in testing for reliability. Please follow the steps carefully and completely:

1. Allocate approximately 2 to 3 hours for the audit tour.
2. Obtain the "double-filled" audit pack which will ensure each observer will have matching questionnaires for the two patients randomly selected by patient type, and a unit observation.
3. Go to a patient care unit and obtain the records for the two patients selected. Pull the appropriate questionnaires for the patients selected.
4. Each observer takes a record and ascertains the information necessary to answer the questions. All observations must be made at the same time by the pair of observers. When recording the answers, observers should not discuss their observations or their responses.
5. Proceed with other sources of information for each questionnaire. When interviewing or observing, go as a team. However only one observer should ask questions (so as not to tax staff or



patients). The other observer should listen (or observe) and record their responses on their own answer sheet - again, without consulting each other. It is important that both observers have the same information collected at the same time for making judgements. The responses to each question should be recorded immediately. Do not wait until leaving the patient's room or area to record answers.

6. Use the same procedure for answering questions on the Unit Specific questionnaire.
7. When the tour is complete, return to Mrs. Meilicke's office where percentage of agreement will be determined and a discussion of the tour will take place.

Thank you.

**APPENDIX J**  
**SAMPLE QUESTIONNAIRE AND ANSWER SHEET**

APPENDIX J

QUESTIONNAIRE 531 FOR QUALITY MONITORING

24JUN83

COMPILED BY  
NURSING AUDIT

## APPENDIX J

## SOURCE OF INFORMATION: PATIENT RECORD

01 1.104

1. ARE DESCRIPTIONS INDICATIVE OF MENTAL-EMOTIONAL STATE RECORDED AT THE TIME OF ADMISSION TO THIS UNIT?

Do not code NA for adults or children; may code NA for infants.

Applies to statements of behavior, e.g., alert, talkative, crying, laughing; or to statements of mental emotional state; e.g., anxious, depressed, mentally retarded, unconscious, not responding.

Code Yes only if statement recorded within first 24 hours of admission.

In Emergency: Code Yes only if statement is recorded prior to observation. May be recorded by either nursing or other health team members for a Yes answer.

In L & D: Code Yes only if statement is recorded prior to the observation.

Source of Information: Patient Record

No: 1      Yes: 2      Not Applicable: 3

01 1.203

2. IS HEIGHT RECORDED UPON ADMISSION TO THIS UNIT?

Code NA if information recorded on admission to another unit.

Code Yes only if information is present and is recorded within 24 hours of admission.

In L & D: Code Yes only if information is present and is recorded prior to the observation.

In Psychiatry: Code Yes only if information is present and is recorded within three days of admission.

Source of Information: Patient Record

No: 1      Yes: 2      Not Applicable: 3

01 1.301

3. IS THERE A WRITTEN STATEMENT ABOUT THE CURRENT CONDITION OF THE SKIN?

Relates to dryness, turgor-hydration, absence or presence of skin lesions, localized skin color, warmth, etc. Do not accept general description such as "pale". Should apply to present status or within past 48 hours.

Code NA only if skin condition is not a real or potential problem.

QUALITY ASSURANCE - ANSWER SHEET

HOSP. #   /1-2      QUEST. #    /3-4-5      AUDITOR #    /6-7-8  
 PT. AGE   /9-10      PT. SEX  M-1 /11      UNIT CODE   /12-13  
 MONTH   /14-15      DAY   /16-17      YEAR   /18-19  
 SHIFT  D-1 /20  
            E-2

NOTE: Circle ONE number for each response

QUES. #	ANSWER	QUES. #	ANSWER
1	1 2 3 4 5 6 7 /25	26	1 2 3 4 5 6 7 /50
2	1 2 3 4 5 6 7 /26	27	1 2 3 4 5 6 7 /51
3	1 2 3 4 5 6 7 /27	28	1 2 3 4 5 6 7 /52
4	1 2 3 4 5 6 7 /28	29	1 2 3 4 5 6 7 /53
5	1 2 3 4 5 6 7 /29	30	1 2 3 4 5 6 7 /54
6	1 2 3 4 5 6 7 /30	31	1 2 3 4 5 6 7 /55
7	1 2 3 4 5 6 7 /31	32	1 2 3 4 5 6 7 /56
8	1 2 3 4 5 6 7 /32	33	1 2 3 4 5 6 7 /57
9	1 2 3 4 5 6 7 /33	34	1 2 3 4 5 6 7 /58
10	1 2 3 4 5 6 7 /34	35	1 2 3 4 5 6 7 /59
11	1 2 3 4 5 6 7 /35	36	1 2 3 4 5 6 7 /60
12	1 2 3 4 5 6 7 /36	37	1 2 3 4 5 6 7 /61
13	1 2 3 4 5 6 7 /37	38	1 2 3 4 5 6 7 /62
14	1 2 3 4 5 6 7 /38	39	1 2 3 4 5 6 7 /63
15	1 2 3 4 5 6 7 /39	40	1 2 3 4 5 6 7 /64
16	1 2 3 4 5 6 7 /40	41	1 2 3 4 5 6 7 /65
17	1 2 3 4 5 6 7 /41	42	1 2 3 4 5 6 7 /66
18	1 2 3 4 5 6 7 /42	43	1 2 3 4 5 6 7 /67
19	1 2 3 4 5 6 7 /43	44	1 2 3 4 5 6 7 /68
20	1 2 3 4 5 6 7 /44	45	1 2 3 4 5 6 7 /69
21	1 2 3 4 5 6 7 /45	46	1 2 3 4 5 6 7 /70
22	1 2 3 4 5 6 7 /46	47	1 2 3 4 5 6 7 /71
23	1 2 3 4 5 6 7 /47	48	1 2 3 4 5 6 7 /72
24	1 2 3 4 5 6 7 /48	49	1 2 3 4 5 6 7 /73
25	1 2 3 4 5 6 7 /49	50	1 2 3 4 5 6 7 /74

\* Check to make sure the EXACT number of questions have been answered. This information is on the audit pack cover. e.g., Questionnaire 511 has 42 questions.

**APPENDIX K**

**PERCENTAGE AGREEMENT FORM AND COMMENTS FORM**

**APPENDIX K**  
**INTEROBSERVER RELIABILITY REPORT**

Date of Audit: \_\_\_\_\_

Station Number: \_\_\_\_\_

Auditor: \_\_\_\_\_

Observer: \_\_\_\_\_

Questionnaire #: \_\_\_\_\_

	Patient One	Patient Two	Unit Observation
Total # of Criteria -			A

Total # of "Agrees" before discussion -			B
---	--	--	---

$\frac{B}{A} =$  \_\_\_\_\_ % (Reliability Score) \*

\*A minimum of 85% agreement necessary for reliable audit

Total # of Criteria -			A
-----------------------	--	--	---

Total # of "Agrees" after discussion -			C
--	--	--	---

$\frac{C}{A} =$  \_\_\_\_\_ % = (Percent of Agreement after Discussion)

**NOTE:** If 100% of agreement is not reached for discussion, make notes below:

QUESTION #: \_\_\_\_\_

QUESTION #: \_\_\_\_\_



Control \_\_\_\_\_

Experimental \_\_\_\_\_

**OBSERVER I**

Position: \_\_\_\_\_

Audit Orientation  
(date): \_\_\_\_\_

Nursing Experience:

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

**OBSERVER II**

Position: \_\_\_\_\_

Audit Orientation  
(date): \_\_\_\_\_

Nursing Experience:

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Criteria No.	Comments	Code

**APPENDIX L**

**SAMPLE LIST OF QUESTIONS AND REASONS FOR DISAGREEMENT**

## APPENDIX L

## SAMPLE LIST OF QUESTIONS AND REASONS FOR DISAGREEMENT

All Sources Combined

Item	Total Number of Occurance	Frequency of Disagreement		Frequency of Agreement	
	f	f	%	f	%
Overall Total	3308	497	15.02	2811	84.98
Reason E: Error in Recording		82	16.50%		
Reason P: Patient Response Unclear		25	5.03%		
Reason N: Nurses Response Unclear		15	3.02%		
Reason Q: Question Unclear		21	4.23%		
Reason S: Standard Varies		12	2.41%		
Reason J: Observer Judgement Varies		143	28.77%		
Reason R: Careless Reading of Question		88	17.71%		
Reason C: Patient Record Confusing		85	17.10%		
Reason O: Other Reason		26	5.23%		

Source 01

Item	Total Number of Occurance	Frequency of Disagreement		Frequency of Agreement	
	f	f	%	f	%
01 4.106A	13	1	7.69	12	92.31
Reason C:	Patient Record Confusing	1		1	100.00%
01 4.106B	13	4	30.77	9	69.23
Reason J:	Observer Judgement Varies	2		2	50.00%
Reason C:	Patient Record Confusing	2		2	50.00%
01 4.106C	13	5	38.46	8	61.54
Reason E:	Error in Recording	2		2	40.00%
Reason J:	Observer Judgement Varies	2		2	40.00%
Reason C:	Patient Record Confusing	1		1	20.00%
01 4.106D	13	1	7.69	12	92.31
Reason C:	Patient Record Confusing	1		1	100.00%
01 4.106E	13	2	15.38	11	84.62
Reason C:	Patient Record Confusing	2		2	100.00%
01 4.111	3	1	33.33	2	66.67
Reason S:	Standard Varies	1		1	100.00%
01 4.201	11	2	18.18	9	81.82
Reason Q:	Question Unclear	1		1	50.00%
Reason R:	Careless Reading of Question	1		1	50.00%
01 4.202	15	4	26.67	11	73.33
Reason R:	Careless Reading of Question	2		2	50.00%
Reason C:	Patient Record Confusing	1		1	25.00%
Reason O:	Other Reason	1		1	25.00%

Source 02

Item	Total Number of Occurance		Frequency of Disagreement		Frequency of Agreement	
	f		f	%	f	%
02 2.907B	18		1	5.56	17	94.44
Reason E:	Error in Recording				1	100.00%
02 2.907C	18		2	11.11	16	88.89
Reason E:	Error in Recording				1	50.00%
Reason Q:	Question Unclear				1	50.00%
02 2.910	4		1	25.00	3	75.00
Reason R:	Careless Reading of Question				1	100.00%
-----						
Objective 2	166		16	9.64	150	90.36
Reason E:	Error in Recording				5	31.25%
Reason Q:	Question Unclear				1	6.25%
Reason J:	Observer Judgement				6	37.50%
Reason R:	Careless Reading of Question				2	12.50%
Reason O	Other Reason				2	12.50%
-----						
Total	166	16	9.64	150	90.36	
Reason E:	Error in Recording				5	31.25%
Reason Q:	Question Unclear				1	6.25%
Reason J:	Observer Judgement-Varies				6	37.50%
Reason R:	Careless Reading of Question				2	12.50%
Reason O:	Other Reason				2	12.50%