THE RELIABILITY AND VALIDITY OF FUNCTIONAL STATUS INDICES USED IN A CLINICAL TRIAL

Susan Jane Boucher

A Thesis Submitted to the
Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements for
the Master's Degree of Health Science
(Rehabilitation)

School of Physical and Occupational Therapy
McGill University Montreal, Québec

Reliability and Validity of Functional Status Indices

Reliability and Validity of Functional Status Indices

ABSTRACT

THE RELIABILITY AND VALIDITY OF FUNCTIONAL STATUS INDICES USED IN A CLINICAL TRIAL

A controlled clinical trial to study the effects of adding a geriatric consultation team to the traditional pattern of care for the elderly patient in an acute care hospital had been conducted. To assess the quality of the data collected on the functional status outcome measures, a supplementary investigation was undertaken. The objectives were to examine the reliability and validity of the Barthel Index and the Level of Rehabilitation Scale (LORS) used in the Trial. Fourteen evaluators and 5 interpreters were trained by the study instructors using videotaped assessments of elderly individuals. Periodic monitorings of the evaluation sessions were conducted in the hospital and home settings. Concurrent Validity was examined through the Functional Status Assessment Instrument (FSAI). Results demonstrated that good to excellent levels of rater reliability were achieved for the duration of the Trial. Furthermore, 75% of the validity coefficients were significant for the three scales.

RESUME

FIABILITÉ ET VALIDITÉ DE DIFFÉRENTS INSTRUMENTS DE MESURE DE L'ÉTAT FONCTIONNEL UTILISÉS DANS UNE ÉTUDE GÉRIATRIQUE

Un essai clinique randomisé a été mené pour étudier les effects de l'addition d'une équipe de consultation en gériatrie au modèle de soins traditionnels aupres des personnes âgées hospitalisées dans un hôpital de soins aigùs. Afin de dèterminer la qualité des données recueillies sur les mesures de l'état fonctionnel, une seconde enquête a ètè entreprise les objectifs étaient d'examiner la fiabilité et la validité de l'Index Barthel et de l'Echelle de niveaux de réhabilitation utilisès dans l'étude. Les responsables de l'étude ont entraîné 14 évaluateurs et 5 interprètes à évaluer les personnes âgées à l'aide de bandes video. Le contrôle pèriodique des sessions d'évaluation s'est effectué en milieu hospitalier et familial. Parallèlement, la validité était vérifiée avec un Instrument de Mesure de l'Etat Fontionnel. Les résultants ont démontré que les niveaux de fiabilité atteints par les évaluateurs s'échelonnaient de bons à excellents pour la durée de l'étude. De plus, 75 percent des coefficients de validité étaient significatifs pour les trois échelles.

PREFACE

Canadians 65 years and older are rapidly becoming an prominent part of today's society and are making up a larger portion of the nation's total population than ever before. In 1901, only 5.0% of the population was 65 and over, however by the year 1983, this proportion had risen to 10% which represented better than 2.5 million people (Statistics Canada, 1984). This trend is predicted to continue, and is estimated to produce a population of 3 to 3.5 million elderly people by the year 2001 or between 11% and 13% of the Canadian populace. However, Canada is not alone. Industrialized countries through out the world are now being confronted with the reality of aging societies. Presently, the over 65 age-groups make up 13.6% of the inhabitants of France, 14.2% of the people of the United Kingdom, 15.1% of the Swedish public, and 10.7% of the population of the United States. In comparsion, Canada is a much younger nation but as previously stated is maturing quickly.

Aging is a complex sequence of biosocial changes (Bromley, 1966). Older people are frequently faced with multiple health problems and increasing disability, rendering them dependent on others and the medical system. Often limited by fixed incomes and in many instances lower standards of living, their problems are fast becoming the nation's concern. An obvious result of today's aging population is the development of a society in which a growing number of older people will become progressively more dependent on the

decreasing proportion of younger people.

One outcome of this phenomena is the rising health care costs required to service this older population. Their impact on the demand for health care services is tremendous and will steadly increase over the next several decades. The elderly person and, in particular, the older elderly accounted for 35% of all hospital patient days in 197[illegible] and are projected to account for 42.5% of patient-days by 200[illegible] (Rombout, 1975[illegible]. In 1973, institutionalized care alone represented the largest element in total health expenditures in Canada, amounting to 4.3 billion dollars or 52% of the total health expenditure of which 3.2 billion was oriented towards acute-care (Rombout, 1975). In the U.S., health care costs for the elderly have risen from 8.2 billion in 1966 to 34.9 billion in 1976 or 29% of the nation's total health care bill, an increase of 190% in real dollars (Kane and Kane, 1978). Similarly in Canada, counting only the Federal disbursement of funds and not the provincial contribution, spending on the elderly in the last fiscal year of 1983-84, came to 17.6 billion dollars, or 18.1% of all Canadian expenditures (Statistics Canada, 1984). Thus it has become apparent that preparing for the projected needs in health care costs for the older person will be an awsome and an ardously expensive endeavor. The stakes will be high and the potential consequences of any misstep will be tremendous.

In sum, the demographic trends for the future, paired with the mounting costs of health care point toward a need for a stronger hand by the nation's policymakers. They must

develop more appropriate and cost-effective plans for providing health care to our present and future elderly populations.

In turn, these policymakers need concrete information on the benefits and costs of various approaches to care for this group. Currently there appears to be few concrete facts to guide policymakers or professionals in the management of the elderly. Various approaches to the care of this segment of the population have been clinically accepted but few have been validated (Bloom and Soper, 1980; Wood Dauphinee and Clarfield, 1984). Thus specific attention must be given to determining the most effective and most efficient means of caring for the older person.

At present, a group of researchers in Montreal are attempting to address one component of this question. At the Royal Victoria Hospital a controlled trial has been conducted to examine the effects of providing coordinated geriatric team care and early rehabilitative efforts for the elderly patient in the acute-care setting. Patients over the age of 70 years admitted to either team care or conventional care have been followed for a period of six months to determine if this geriatric team has been able to effect favourable results for a series of preselected outcome variables.

Whenever a major clinical trial is undertaken, assuring both the adherence to the study protocol and the quality of the data collected becomes mandatory. Knowledge concerning the accuracy and precision of the measurement and of the process and outcome variables is also extremely important. In

examining the data, the sources and extent of systematic variation need to be identified. Equally, the instruments of measurement must be evaluated to determine if the objectives of the measuring tool produce data relative to the purpose of the project under study (Henderson, 1975; Potvin, 1975; Garraway, 1976; and Knatterud, 1981.

In view of these concerns, a second investigation has been undertaken to examine the reliability and validity of the data obtained from this clinical trial of geriatric patients. The major goal of this research is to determine if the geriatric study findings are meaningful. In other words, is the study measuring what it purports to measure and is it obtaining data that are reproducible.

This thesis is organized into six chapters. The first chapter presents a review of pertinent literature related to the care of the geriatric patient and how that care and its outcomes are assessed. Specifically, the first section focuses on the locus of care, the philosophy of care, the care givers and the effectiveness of care and its achievements. The second section of chapter I addresses issues of measurment procedures and the difficulties inherent in these processes.

Chapter II presents the Parent Study. This section includes the description of the facilities at the Royal Victoria Hospital as well as the objectives and design of the main study. The two approaches to care delivery and the methods, procedures and instruments used in the collection of the data are also described.

Chapter III describes the present study of quality assessment. The objectives, hypothesis and methodology are introduced. Specifically this chapter describes the study design and includes the description of the study evaluators, interpretors, and instructors. Measurement indices are described in detail. In addition, the outline to the analysis of the data is presented.

Chapter IV reports the results. This chapter is subdivided into two sections: the reliability study (part I) and the validity study (part II). In part I, a comparison is made between the three groups of raters and the gold standard. Percentage of agreement, measurement bias and the analysis of rater variance is examined. The three groups of raters are then examined for their inter and intra-rater reliability. Part II of the results chapter addresses the issues of establishing the validity of the study's functional scales.

In Chapter V, a discussion of the findings is provided. Each section of the results chapter is considered separately.

To conclude, Chapter VI deals with the summary of the study, the implications of the findings are discussed with respect to future investigators and health care professionals and the limitations of the study are presented.

ACKNOWLEDGEMENTS

Many people have provided me with assistance during the planning and executing of this thesis, but it is Dr. Sharon Wood-Dauphinee, Associate Director of the School of Physical Therapy and my thesis advisor to whom I am most indebted. Over a five year period she has advised, supported and encouraged me to the completion of this task.

Dr. James Hanley, my statistics advisor and teacher, provided me with his time, expertise, and explanations during the phase of data analysis. In addition, I would like to extend my gratitude to Marielle Olivier, programmer with the Department of Epidemiology and Health Care for her astuteness in identifing many taxing computer problems and to Nicola Richards Consultant McGill Computing Centre for her energy and interest in bringing this manuscript to its final stages.

In the clinical setting, I owe a great debt to Dr. David Gayton, Director of the Divison of Geriatrics, Department of Medicine at the Royal Victoria Hospital. He gave me the opportunity to evaluate the quality of the data collected for the Geriatric Trial which was in progress. Many thanks go to Mme Marie de Lorimier, the Study Coordinator, and to the raters of the Trial for their time and patiences during the numerous testing sessions.

I also owe a debt to the administration of the Centre de Réadaptation Constance Lethbridge, to Mrs. Jane Petrov, Librarian for helping me in the video-tape sessions and in particular to the six patients who volunteered to be evaluated

and filmed for this project.

Financial assistance is also gratefully acknowledged. The McGill Graduate Faculty Summer Research Fellowship provided a start up grant and the Fonds de la Recherche en Santé du Québec funded the major portion of the project (Dossier [illegible]).

Finally, [illegible] my husband for his understanding, support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

The Literature Review

Section I

Care of the Elderly Patient

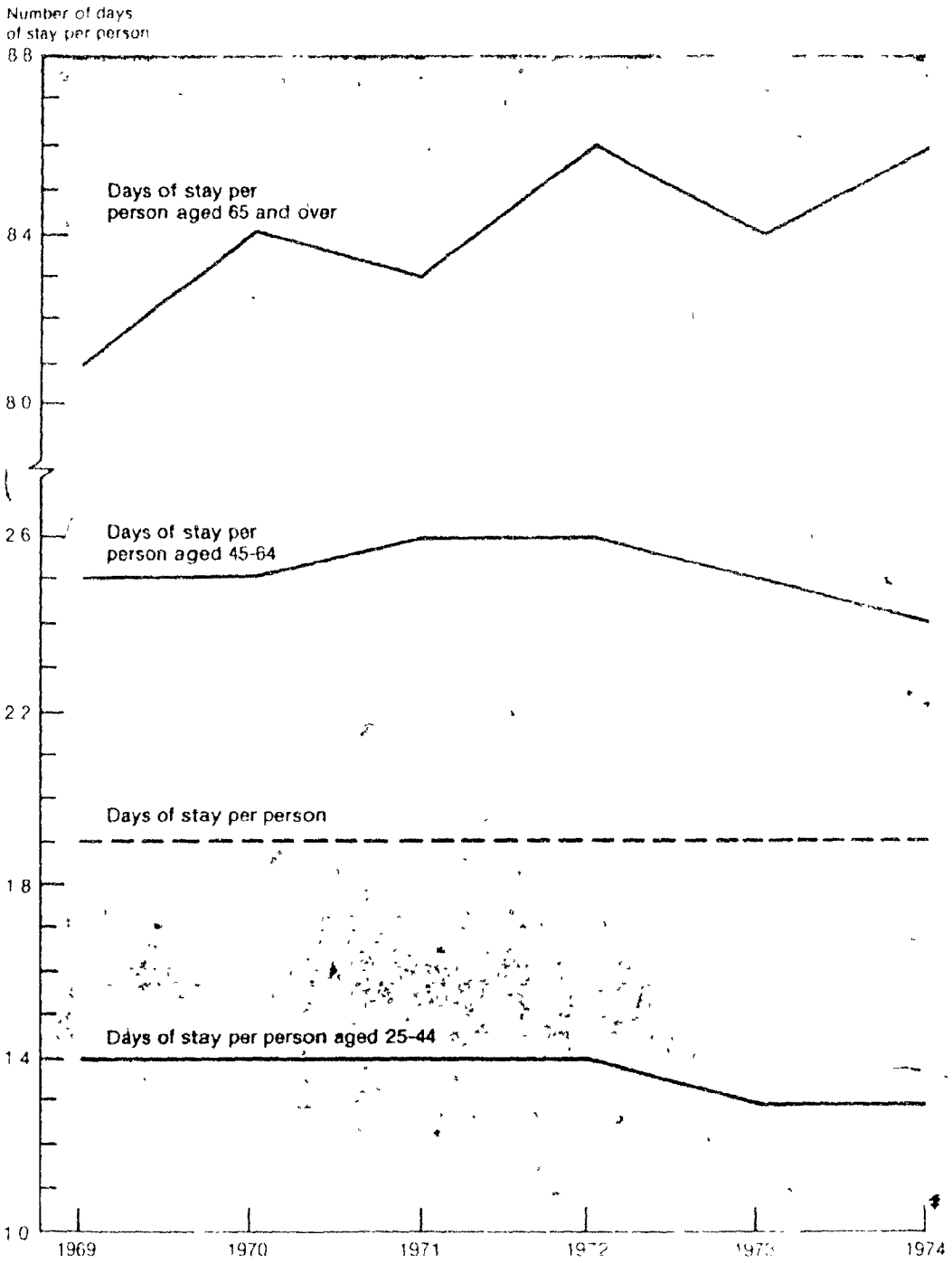
Two major factors contributing to the loss of independence in persons of any age are chronic disease and disability. In 1972, an estimated 1.7% of the institutionalized American citizens of all age groups were listed as being impaired to some degree in the performance of at least one daily activity. Of this populace, half reported difficulties in carrying out the major functions of meal preparation, house-keeping and shopping as a result of health-related problems. Yet, what is most striking is the marked loss of independence amongst the elderly population, with an estimated seven million people falling into these categories in the United States alone (O'Brien, 1982). Although similar comparisons could not be found for Canada, it is possible to imagine comparable figures for this country.

Clearly people are living longer today in contrast to fifty years ago, however, longevity is not necessarily synonymous with good health. As Katz (1983) pointed out, the prevalence of chronic illness has been most evident amongst the older generations, with the older elderly (75+) averaging

three or four concurrent chronic illnesses. This increase in chronic illness and longer life expectancy has had a direct impact on hospital and nursing home stays. By definition, length of stay is the number of days a patient spends in the hospital. In general this overall length of stay has gradually fallen in Canada, as seen in Figure 1-1, with one exception, patients 65 and over.

Figure 1-1

Average Number of Days of Hospital Stay per Person for Selected Age Groups, Canada, 1969 to 1974



Source: Stone, Leroy O. and Fletcher, Susan. 1980. A Profile of Canada's Older Population. Montréal: Institute for Research on Public Policy.

It is therefore, not suprising that chronic disease is rapidly becoming the most frequently encountered problem in medicine today. Moreover, these disorders frequently lead to impairment of function with resultant disabilities. An elderly person with a disability may need assistance in bathing or household activities. This need for help is most often recognized when there has been a loss of an able relative or friend. The problem may be further compounded if the older people lacks some form of social interaction and thus may fall victim to social isolation. It has been stated that retirement is often associated with decreased productivity. Financial resources begin to deplete just as health care needs and risk which can result in additional serious consequences for the person, family and community

The elderly segment of the population is growing at an astonishing rate. Of this group, a greater number of people are subject to the common disorders of old age such as cardiovascular diseases, cancer, arthritis, and fractures. Faced with this increased survival rate of people after 65 and the health-related problems that are likely to be present, it is clear that the health-care mechanisms must be addressed on a long term basis.

Long-term care is multifacted. It is characterized by myriad requirements, multiple services, family involvement, community institutions as well as the presence of government support (Vogel and Palmer, 1983). Available information suggests that people who need long-term care are likely to

have two or more illnesses that require attention from a multitude of services. In addition, older people frequently experience repeated hospitalizations. As a result, there are numerous transfers from one level of care to another (Katz, 1983). Meanwhile, the costs of this care continue to escalate.

There is no cure for chronic disease. Instead the goals for long term care are restoration and maintenance of function at a maximum level for that particular individual. This long-term care could be in the form of community services which would enable the older individual to remain at home. Through this approach informal help systems could be organized through family, friends, and neighbors and, long term care could mean institutionalization. Some long term care may be in a treatment situation and in longer term care where multidisciplinary services need to be integrated with basic living supports (Katz, 1983).

Locus of Care

Community Services

Within the continuum of long-term care, adult day care planning is an area that is slowly achieving recognition. Presently, adult day care is an umbrella term which blends psychosocial and health services in varying intensities and is provided in a variety of institutional setting. Adult day

care has been defined as a program of services provided in an ambulatory setting for adults who do not require 24 hour institutional care but because of physical and mental disabilities are not able to live independently on a full-time basis (O'Brien, 1982). However, there is no one definition of adult day-care and similarly, there can be no one model.

In order to be a viable structure, adult day-care programs need to identify the existing community resources and to amalgamate with them to form a network of comprehensive services for their elderly populace.

In concordance with this concept, an interconnecting chain of services can be offered to the individual, in the form of health care supervision and essential support activities of daily living. This serves the person to continue to function as an active member within the neighbourhood, moving freely in and out of the various community service settings, as is required. Yet, for when the individual's health demands further attention, a direct liason with more intensive therapies can be provided. Within this context, adult day care programming is a part of the long-term care system but it is not understood as an alternative to institutionalization (O'Brien, 1982).

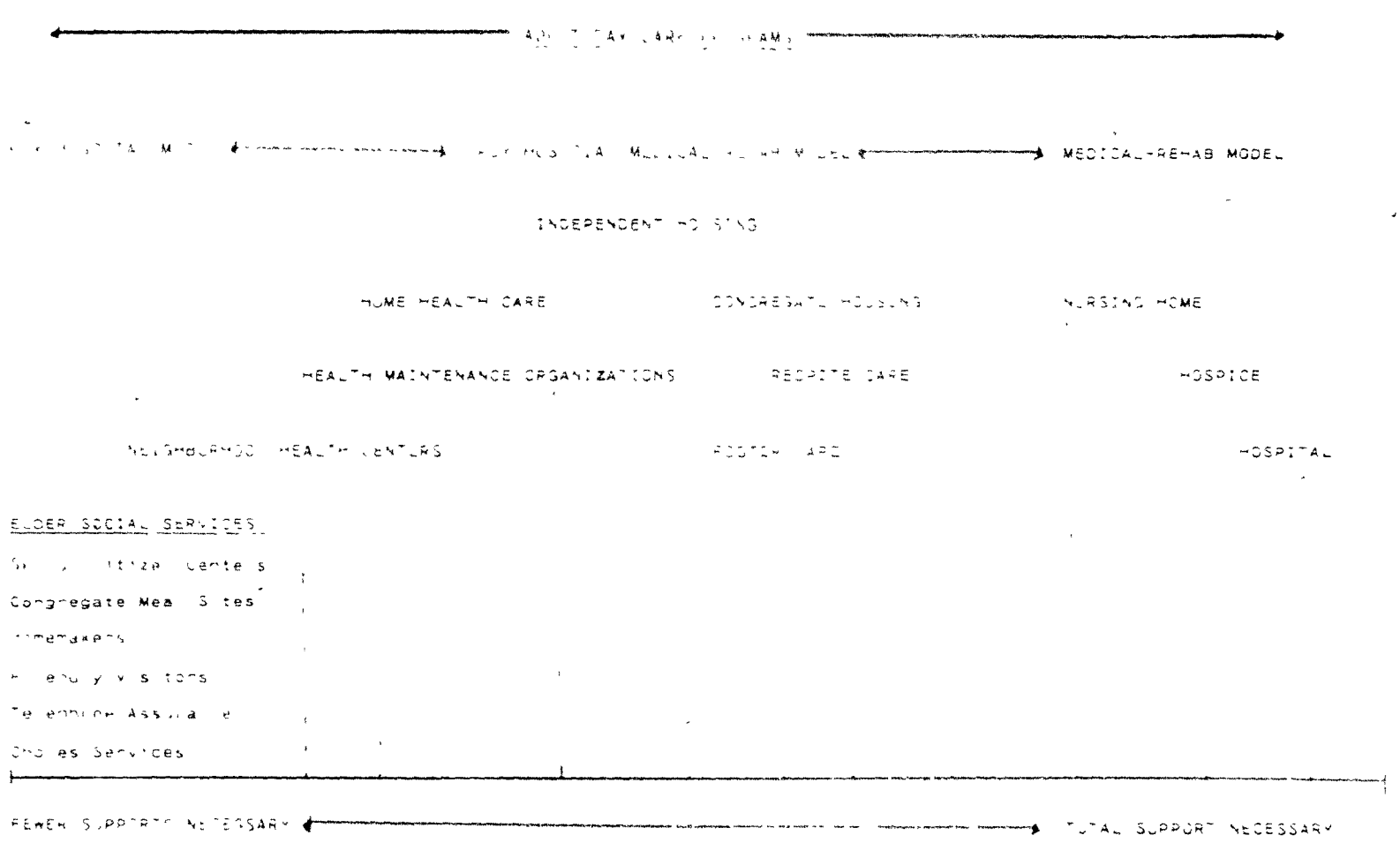


Figure 2 Model of a Continuum of Services for the Elderly
Adapted from Planning for Adult Day Care by D. Brody (1977)

As seen in the Figure 1-2, (Clark 1982), there are numerous components which can make up the continuum of community services for the elderly. These elements can extend from the lowest level of support or care, through to providing total care and support to the elderly and their families. Looking at the left side of the continuum, the more independent individual may require only the socialization support available through friendly visitors, congregate meal sites or senior citizen centers. Further along the model, the protective environment of independent housing or a foster homes for the elderly can provide the individual with sufficient support to remain in the community. While at the other extreme, the intensive services from the rehabilitation center, the nursing home, respite and hospice centers, or the acute-care hospital permits the individual to regain certain levels of functional status and thus some degree of independence.

At several points along the continuum, resources can overlap. This overlapping of services allows the individual to receive intensive management when needed and the possiblity of returning to a reduce level of assistance when and if health permits. In effect, the person is given the option of choice.

Hospitals

Despite the various alternatives to care, the hospital as an institution nevertheless, remains a central point in the lives of many elderly people. While it has been established that community services may provide this elderly population with a means of staying in their homes; an increasing number of elderly with acute, subacute, rehabilitative, and chronic problems are still finding their way into the costly institutions.

A hospital can be classified as an acute or a long term care facility. The acute care hospital provides full emergency care and a variety medical and surgical services to the community. Whereas, the long term or chronic care hospital stresses restorative and maintenance care rather than primary treatment of disease. In between these two facilities, a third type of institution, the rehabilitation center places emphasis on progressive rehabilitative techniques for those individuals with a good prognosis for recovery.

Clearly, one thing that each of these facilities have in common is the rising numbers of older people seeking their assistance. As previously described, many elderly individuals have a multiplicity of chronic diseases leading to either hospitalization or institutionalization. In consequence, the average length of stay and the total number of hospital days per year escalate in relation to the age of the subject, with a steep rise in the 75 years and older age group Lamort et

al.,1983).

During 1979, the average length of hospital stay in the United States amounted to 11.4 days for the older elderly aged 75 and over and 10.1 days for younger elderly (65 to 74). In comparison, the middle-aged (45 to 64) recorded 8.2 days with the shortest stay of 7.2 days reported for all other ages. The older elderly used 6062 acute-care hospital days per year per 1000 population, compared with 3124 days for the younger elderly. By far, these figures have more than tripled those reported for all other age groups (Lamert et al., 1983).

Similar findings have been reported for a Canadian population in a study by Robertson and Rockwood (1982). In describing 828 consecutive admissions of older patients to three hospitals in the city of Saskatoon, they were also able to demonstrate that hospital admissions and average length of stay increased with advancing years. By 1976, Canadian people over 65 made up only 8.6% of the population yet used 38% of the patient days in general hospitals. The average hospital stay for all ages was 10 days for men and 11 days for women while men over 75 years averaged 27 days and older women 35 days and this trend was projected to continue.

As Brody 1976 suggested the acute-care hospital will continue to be a focal point in the lives of the older population. Although the elderly growth factor is a major contributor to the higher occupancy rate of older patients in the acute care hospital, other important factors need to be delineated. In some instance, there may be inappropriate

admissions of patients to the hospital due to lack of family support. Once admitted, there maybe lack of coordinated services aimed at improving functional recovery, coupled with the lack of discharge planning and the wait of facilities for appropriate follow-up care. Moreover, there is a shortage of less intensive modes of care such as home care, day care or nursing homes. In support of this evidence, it is essential that the varied needs of the older person be carefully considered. In this time of limited monetary resources and budgetary constraints, it is crucial that health care services be reorganized with emphasis on improving the functional independence of the elderly patient. One means of achieving this goal may be active treatment programs specially designed for the elderly. Through these specialized services fewer admissions, reduced hospital stay and reduced health care costs may be achieved. Obviously, the argument for the restructuring of health care services for the elderly is compelling.

Specialized Geriatric Care Units

There has been an impetus to develop an intervention mechanism to meet the demands of chronic illness among the elderly (Brody, 1976). In 1978, Kane and Kane drew further attention to the necessity for more effective alternatives to long-term care for the older patient. Through the use of specific health-care programs, they stated that many people who were currently being institutionalized could be cared for

without resorting to chronic wards and eventually to nursing homes. One advancement towards their attainment of this goal was to streamline the existing hospital care offered to the elderly patient. This entailed the creation of specialized geographic units which provided comprehensive evaluation and treatment for the multiple needs of the older person.

Traditionally, Great Britain and Europe have been the trend setters in the refinement of health care services and this innovation was no exception. Marjorie Warren, one of the founders of modern geriatric medicine was instrumental in initiating this concept of specialized geriatric assessment units during the late 1930's. As a result of her experience, Dr. Warren advocated the use of comprehensive assessments and an attempt at rehabilitation for all elderly patients before admitting them to long term hospitals. Her pioneering approach has remained a basic principle of British gerontology (Rubenstein 1984).

Early descriptions of the patients admitted to geriatric units and the treatment provided have frequently been recorded in the literature (Exton Smith 1962; Silver and Outerbridge (1965); Hodkins and Jeffreys (1972); Barrick et al.,(1976); Hodkinson and Hodkinson (1980)). In general, these studies concluded that the majority of older patients participating in the program had benefited from the specialized geriatric services offered. The elderly person definitely required intensive therapy but only for a comparatively short period of time, which averaged between 60 to 90 days.

Through the use of geriatric units, Hodkinson and Jeffreys (1972) took the concept one step further by suggesting that the ill elderly person could be offered a different approach to care than was traditionally provided in the usual hospital ward. Working in a geriatric service of a new district general hospital, these authors encouraged local practitioners to make direct referrals of patients to the service. No preadmission assessments of patients were necessary, thereby facilitating easier access. The emphasis of the unit was on active treatment and rehabilitation with an early discharge policy aimed at avoiding the development of a working ward. It was the belief that acutely ill patients were prone to deterioration the longer they remained in the hospital environment and further complications. Within the first four weeks of admission, 48% of patients were discharged, however of those remaining only 31% were discharged within the next four week period. Furthermore, mortality demonstrated a 19% risk of death within the first month of admission but had dropped to 8% by the second month. After three months they reported that discharges home became increasingly difficult due to institutionalization and social dependence as the patient's ties with the community began to wither. The major problem in the field of geriatrics, they felt, was that of "underexpectation" which affected the patients, the public, and most importantly the medical profession itself. In response, the authors advocated that with active policies and adequate and enthusiastic staff, geriatric services could be revolutionized, resulting in

higher turnover of patients, reduction of hospital stay and no waiting lists.

Hodkinson et al. (1980), followed with a second study from which they concluded that the likehood of death in the hospital was markedly increased for the older elderly groups and the percentage of discharges home were limited. Other factors adversely affecting discharge were low mental test scores, unnecessary waiting for admission to the hospital and the previous inactivity of the individual.

O'Brien et al. (1973), confirmed these findings, while studying a geriatric department in an acute care setting where patient turnover more than kept pace with demand. The authors postulated that the patients had come to see themselves as [illegible] treated even when they were [illegible] irremediable.

In 1966, Sloane examined the dilemma of the primary physicians faced with the decision whether to admit the older person to a general hospital or to place them directly into a nursing home. With a small sample of 29 subjects destined to be eventually placed in a nurse home, he was able to demonstrate that eight of the 29 patients benefited from hospitalization when comprehensive assessments and treatments were available. The outcome was community placement for these eight people who now required less intensive care than that provided in a nursing home setting. He further emphasised that functional independence in ADL, reasonable mental function, and the availability of family support on discharge were important factors in predicting those patients who would

benefit most from intensive rehabilitation and those who should be considered inevitable recipients of long-term care. In conclusion, the author warned that if patients were transferred from a geriatric unit to a nursing home as an intermediately care step, they would most likely remain there despite the regaining of sufficient independence to live successfully in a less protective environment.

In reflection, now is the time for health-care planners to seriously consider the health needs and approaches to care for the elderly population. The older patients are marked by numerous medical and social problems, and if not properly monitored, these conditions hold the potential for substantial degrees of disability and recurrent hospitalizations. Furthermore, it also has become increasingly important to control the overloading of limited acute care resources with long term chronic cases. In fact, the word "acute" as it applies to general hospital beds for the elderly patient, needs to be re-evaluated. Upon examination, many questions still remain unanswered. Nevertheless, as O'Brien and colleagues (1973), have suggested it is only common sense that every hospital group earnestly consider the development of specialized geriatric services.

Philosophy of Geriatric Team Care

Over the past 25 years, the concepts of teams, team care and comprehensive care have gradually taken a place in our

working terminology. Ideas of team work can be found in scientific, professional and commerical enterprises, and in particular in the field of medicine. As Rothberg (1981), indicated the idea arose from the need to deal with the increasingly complicated delivery of health services that resulted from the knowledge explosion in basic science and medical technology. The team concept of health care delivery evolved as a compromise between the benefits of specialization and the need for continuity and comprehensiveness of care.

In 1976, Halstead, in a review of the literature, defined team care as "coordinated, comprehensive care provided by persons who integrated their observations, expertise and decisions", in short, a gestalt (Campbell, 1981). Initially, the specialities of psychiatry, rehabilitation and primary care were the strong supporters of this team concept but gerontology was soon to follow. Clarfield (1982) described comprehensive teamwork as the backbone of geriatric care. He stated that the team has become a means by which diagnosis, treatment and care can be provided in a coordinated manner by specialists representing several disciplines. As part of this approach, health practitioners are constantly encouraged to consider the whole person and nowhere else is this concept more relevant than in the care of our elderly population.

Older persons are complex individuals needing specific but more broadly based and interdisciplinary approaches to their care. If neglected, a genuine danger can exist, for in trying to meet the very real health needs of the aged, we can exacerbate the social and psychological problems of a group

already vulnerable because of losses associated with the aging processes (Kane and Kane, 1978). These patients' needs go beyond the medical scope. To be precise, life styles have to be considered within the range and boundaries that illness and its chronicity allows.

In support of this philosophy, Lefton (1979), stated that the elderly patient's care involved more than the remission or the control of disease. It implied a conception of treatment that views success in terms of the quality of a person life in addition to the conditions of a person's health.

This is the approach a geriatric team tempts to offer. By going beyond the medical needs of the patient, the concepts and theories of the social, behavioral and environmental needs of the individual are also considered and can result in older individuals maintaining an active role with in the community (Lefton, 1979).

It is becoming increasingly evident that geriatrics has now become synonymous with the geriatric assessment unit and the team (Lefton, 1979; Kane and Kane, 1978; Rubenstein, 1981). Specialized geriatric assessment units have been appearing across North America in response to the growing recognition of the many unmet needs of the frail older person and the conviction that these units could have major beneficial impacts (Schuman et al.,1978; Cheah et al., 1979; Checkryn and Roos 1979; Rubenstein et al., 1981; Clarfield 1982; Applegate et al., 1982; Campion et al., 1983; Lefton et al.,1983; and Lichtenstein and Winograd 1984 . The philosophy and organization of the these units has varied according to

the objectives of the center concerned but generally most have included interdisciplinary teams which focus on comprehensive assessment, treatment and rehabilitation for medical, functional and psychosocial problems.

It is now understood that the older person is subject to multiple diagnoses. His physical, medical and social well-being are very closely interrelated, so that multidimensional evaluations of health status are necessary. Measures of functional status that examine the physical and mental disability, the ability to function independently despite disease, and social deprivation of the individual are the most useful overall indicators which assist those who care for the elderly (Kane and Kane, 1981). Geriatricians in Great Britian suggest that rate of change in functioning may be an important diagnostic and prognostic tool.

One of the first papers to relate the functions of the geriatric team within an elderly assessment unit was a descriptive study authored by Schuman et al., (1978). Through a hospital for the chronically ill, he described the impact of this new geriatric program. After the first year of operation, he concluded that treatments actively organized through team planning were able to produce positive results, noting that the majority of their patients improved in many activities of daily living. In addition, he claimed greater efficiency in bed use was achieved by a higher turnover rate owing to the larger number of discharges after a shorter overall stay.

Rubenstein and colleagues (1981), followed with an account of a geriatric evaluation unit at a Veterns Medical Center staffed by a full-time interdisciplinary team. Though they felt improved placement was the most dramatic positive outcome, indentification of many new diagnoses in patients previously evaluated in a general hospital, and the reduction of prescription drugs also contributed to the overall improvement in care.

Another major problem which effects the delivery of health care to the elderly is the lack of appropriate instruction within the medical education of the North American health professions. Clarfield (1982), emphasized this point in describing the workings of a geriatric team in an acute care hospital. One of the primary purposes of this geriatric unit was to form an environment in which medical students, interns and family practice residents could learn geriatrics and team participation. He felt that without a true team effort it was not possible to run a geriatric unit effectively. Although satisfied with the progress of the team, he noted that difficulties existed in offering suitable, conservative and humane medical care to an older population in an agressive acute care setting. In concluding, he stressed that general hospitals would continue to be confronted with an increasingly large number of elderly patients and that the geriatric team was one means of dealing with this difficult situation.

Other workers have further supported Clarfield's point of view. Blumfield and coworkers (1982), promoted the belief

that improvement in approaches and skills of staff through an educational process would improve patient outcomes and decrease length of stay. In citing the 1978 Report on Aging and Medical Education from the Institute of Medicine, she reiterated the need for more formal education in geriatrics for all health professionals. To meet the challenge, a geriatric evaluation team was established to educate and treat elderly individuals. In the author's estimation this team was successful in exposing all professionals to the importance of assessing the needs of the older patient as well as identifying community resources. In turn the elderly patient benefited from increased attention, new diagnoses were made and treated because of the interventions of the team. She concluded by stating that the geriatric team would continue to be an educational and consultation modality for professionals at the site of their interaction with older patients and at the times of crises when such inputs were most relevant.

Other advocates of the team approach (Applegate et al., 1983; Campion et al., 1983; Lefton et al., 1983; and Lichtenstein et al., 1984) have reported Geriatric Assessment Units to be conducive maximizing functional gains in the elderly. In some instances, patients previously slated for institutionalization were redirected to lower levels of care upon discharge from the hospital. Kane and Kane (1981), also noted that multidimensional assessment measures could be used as prediction tools to determine which patients were more likely to remain in the community. In support of this, these scales could served to establish norms in determining specific

places of residence for the elderly population. Other authors have claimed that these units increase the awareness of the special needs of the older individual and produce an enthusiastic and concerted effort from the team to improve the quality of the medical care provided. On a similar theme, Rubenstein (1983), raised an important point in which he emphasised that assessment programs could provide essential data for the identification of a subgroups of patients who could be expected to maximally benefit from these programs. In this manner, the unit could make more efficient use of the scarce health-care resources. Finally, Applegate et al., (1984) and Lichtenstein et al., (1984) contended that rehabilitation for the elderly was a powerful therapy, demonstrating a potential to dramatically effect the disposition of the older person upon discharge from the hospital.

All things considered, each study recognized the increasing need to evaluate geriatric health care within a comprehensive network. In summary, Brody and coworkers (1976), captured the feeling of many fellow researchers when they stated that the creation and utilization of a diagnosis and treatment center for the aged is seminal to major changes in the organization of the short term acute hospital.

Effectiveness of Care

The use of an interdisciplinary approach to health care for the elderly has now reached a level of clinical acceptance. The literature reveals an increasing number of detailed descriptive studies which delineate the establishment of geriatric teams and units. These geriatric specialists have set forth a number of important outcomes measures of hospitalization for acute illness in the elderly person (Brody et al., 1976; Henrickson et al., 1976; Schuman et al., 1981; Palmer et al., 1982; Greenberg et al., 1982; Applegate et al., 1984; Lamont et al., 1983; Fatty, 1983; Patterson et al., 1984). Frequently, functional status, placement, diagnostic accuracy and drug reduction have been proposed as measures of patient outcomes.

Nevertheless, as Halstead (1976) reported, firm proof for the effectiveness of this team care approach is largely lacking despite the existence of a few well-designed studies. In general, most investigations have taken the descriptive or quasi-experimental format which are not designed to fully establish the effectiveness of study interventions.

A descriptive study is described as a survey which aims to investigate the characteristics of a specific population (Abramson, 1979). Such a study does not have a comparison group and cannot demonstrate cause and effect (Applegate et al. 1985). Alternatively, a quasi-experimental design is defined as a study which generally lacks the full control over the scheduling of experimental stimuli (Campbell and Stanley,

1963). There are many reasons why a study may fall short of being a true experiment. The investigator may not have the power to decide who will be exposed to or excluded from the factors under study. There may be no comparison group or no assurance that the experimental and control groups are similar (Abramson, 1979). Nevertheless, even though true experimental situations are not always feasable in the field of health care, some form of scientific investigation into program development needs to considered to derive better understanding of their effectiveness.

A few researchers have used quasi experimental designs with comparison groups to examine the differences in study outcome variables. One such study, (Schuman et al., 1978) examined the impact of a new geriatric program in a hospital for the chronically ill. The control group was selected from medical charts for all patients discharged during the year prior to the new program. Findings indicated that during the geriatric evaluation program, the mean length of hospital stay decreased and discharges home increased when the study patients were compared to those treated prior to the establishment of the program.

In another study (Leftor et al., 1983), patients over 70 years admitted to the Medical Unit for the Elderly were compared to matched controls excluded from the unit due to lack of bed availability. The results indicated that patients from this new Unit were more frequently discharged to the home setting. In addition, they tended to be more independent in activities of daily living, ambulation and mental functioning.

Follow-up evaluations of both groups demonstrated that the place of residence upon discharge had not changed significantly over a three to six month period. While experimental subjects maintained the functional status achieved on discharge their counterparts showed continued improvement in independence.

Other researchers (Campion et al., 1983) described a controlled trial of a geriatric consultation service in an acute care setting and assessed its effects on the care provided to the elderly patients and on the subsequent outcomes. Comparisons were drawn with elderly patients from two similar units. Results indicated that the experimental subjects markedly used more rehabilitation services than their control counterparts, without an increase in length of stay. There was, however, no difference among the three groups in decreasing rates of readmissions over a ten and one-half month follow-up period. These authors concluded that specific geriatric consultation services promoted a better understanding of the field of geriatrics; taught interdisciplinary teamwork; and improved awareness of functional problems of the patient but had to admit that the desired outcomes were not clearly evident. Further, Campion et al., (1983) pointed out that the inpatient geriatric unit strategy entailed far more comprehensive control of management than was possible by simply a consultation team. In addition, this author felt that to be effective, interventions must be more longitudinal and more community-based. Earlier, Burley et al., (1979) proposed a similar philosphy emphasising that

the control over after care was a reason for the success of a British geriatric service in reducing length of stay and increasing the proportion of discharges home.

Other investigators in the state of Texas, (Teasdale, 1983 reported a similar study which had been conducted to examine whether the care in a geriatric assessment unit had any impact on patient placement outcomes when compared to a general medicine cohort. These authors declared that a difference could be detected between the two groups of patients in terms of placement at discharge from initial experience. Further, the mean length of stay for the geriatric unit was 41 hospitals days where as the general medicine unit was only 15 days. In addition, it was suggested that preadmission selection criteria may be crucial to the demonstration of significantly increased frequency of home placement.

Rubenstein and Kane (1984), responded to Teasdale's challenge by stressing the importance of targeting the geriatric assessment activities to a more appropriate population of patients who would likely benefit from the comprehensive care. In their experience only 15 to 20% of elderly patients admitted to acute care hospitals derived better outcomes from being on a geriatric assessment unit than from being on an acute care ward alone. This subgroup included those too frail to return home following their acute-care stay, and those with too poor a prognosis to derive major benefit from the program.

Johnson (1984), also responded to Teasdale's group by

arguing that patient characteristics need to be adequately described so that there are no unanswered questions as to the type of patients studied. He emphasised the importance of determining the degree of disability for the participating subjects, of listing the geriatric diagnoses, and determining the degree of comparability of the study groups. In closing, he stressed that geriatrics should not be viewed as an area of health care characterized by either a clinical state and the presence or absence of an acute illness. Rather, he emphasised the examination of treatment such as altered physiology, advanced age, the presence of multiple chronic diseases, functional impairments, and the role of psychosocial factors in geriatric practice [illegible].

In sum, much has been assimilated from these studies. Nevertheless, the literature on research methodology indicates that nonrandomly assigned control groups contain numerous threats to the internal validity of a study and thus are of limited value in demonstrating effectiveness (Rubenstein et al., 1981; Campbell and Stanley, 1973; and Feinstein, 1977). Even though many patients demonstrated improvement, these afore mentioned studies did not adequately remove the inherent problems due to the lack of randomization. Furthermore, it was not clearly established whether the observed outcomes were due to factors under consideration or differentiation between the groups. By definition, randomization means that patients are allocated to either a treatment group or a control group by chance. Within the

limits of chance variation, randomization should make the control and experimental groups similar at the start of an investigation and ensure that personal judgement and prejudices of the investigators do not influence allocation (Last, 1983). Through this procedure the effects of extraneous study factors can be reduced, and therefore, provide a basis for assessing error.

In contrast to the majority of studies cited thus far in the geriatric evaluation literature, Rubenstein et al. (1984) randomly assigned patients with high probabilities of nursing home placement to their geriatric evaluation unit or a control group. One year later, the investigators demonstrated lower mortality rates for those patients assigned to the specialized unit (23.8 vs 48.4%). They stated that the experimental group was less likely to be discharged to a nursing home initially (12.7 vs 30%) or to be a resident of a nursing home at any time during the follow up period of two years (26.9 vs 46.7%). Furthermore, the authors reported that the experimental patients were significantly more likely to have improvement in functional status and morale than their counterparts. The conclusions were that an appropriate group of patients could benefit substantially from the specialized geriatric unit. However, they cautioned against drawing generalizations from this work as the study was conducted in a atypical environment.

Nevertheless Rubenstein and colleagues (1984 have

presented a carefully designed study which will set the way for future work in the field of gerontology. The challenge now is to reproduce these findings in other settings. As health care costs continue to escalate, priorities must be established by the health service planners when the creation of new programs and the continuation of existing services are reconsidered. Before costly resources are designated to meet these perceived needs, health care data must be collected and analyzed to determine which alternatives will be [illegible]

Section II

Measuring and Assessing the Quality of Care

A major task of clinical research involves the development of methods to monitor the performance of the health care delivery system. An important part of this work encompasses the problem of measuring and assessing the quality of health care. With the coming of the Professional Standards Review Organization (P.S.R.O) in 1972 and the Health Maintenance Organization of 1973, researchers in the United States were confronted with the problem of determining how best to measure quality of care. First introduced by Donabedian (1966), the concepts of quality assessment rely on three basic types of information concerning health care: the effects of care "outcome evaluation", the performance of activities, "process evaluation", and the facilities and settings, "structure evaluation" (Abramson, 1979). By definition, "outcome" refers to what happens to the patient in terms of changes in health; "process", to what a provider does to and for the patient; and "structure", to innate characteristics of personnel or facilities (Brooks et al., 1977).

If one examines the field of quality assessment, differences of opinion exist as to whether process, structure or outcome studies should be considered separately or in combination. In the past, most efforts to assess quality of care have concentrated on the structure and process mechanisms

(Anderson, 1969; and Brooks, 1973). This was because of the general assumption that adequate resources and technology, (structure) contributed to adequate diagnostic assessments and treatment, (process) which in turn resulted in favorable health status (outcome) (Brooks et al., 1977). The structural mechanisms were chiefly concerned with the descriptive, innate characteristics of the facilities and its human resources. While the process methods were concerned with what the physician did to and for the patient, thus measuring the "technical" aspects of care. Because the process evaluation relied heavily on the medical record for data, questions were raised as the to accuracy of these records and if they truly reflected what actually happened during the delivery of care. As a result, the usefulness of measuring the quality of medical care through the evaluation of resources and treatment procedures began to be disputed. The fact was the relationship between the medical care process and the health status of the individual was not always direct. Several studies reported that favorable outcomes had been achieved in the presence of poor process mechanisms. While in contrast, adequate process mechanisms were shown to be confounded by intervening variables, and therefore, failed to demonstrate the desired outcomes. In sum, the validity for using "structure" and "process" procedures to assess medical care was brought into question (Brooks, 1973; Fessel et al., 1972; Nobrega et al., 1977; Romm et al., 1976).

Gradually, public policy began to take a new direction. Outcome measures were considered to be the more valid for

purposes of quality assessment, in recognition that the goal of medical care was to maintain and improve health status. Focusing on the outcome measures, the patient himself becames the source of information. However, because of feasibility problems in the measurement of the long-term outcomes, attention was directed toward the short-term "proximate" outcomes, including both physical and psychosocial factors (Brooks et al., 1977).

However, not everyone supported the use of outcome measures to evaluate the quality of care. McAuliffe (1979), defied the widespread view that outcome variables were superior to process measures for assessing quality of care. As he stated, analysis showed that there were similar sets of problems encountered whether one measured process or outcome.

Faced with these indecisions, several authors have attempted to assess the combined process-outcome relationship and reported mixed results. Support for this association was achieved by using distinct process measures which possessed a conceptual affiliation with the studied outcome variables (Starfield and Scheff, 1972; Langer and Rodin, 1976; Greenfield et al., 1981). It was felt that the knowledge of the process of care could provide assistance in understanding the outcomes achieved.

Conversely, a variety of other studies challenged this interplay of process and outcome measures and were able to demonstrate that in fact no relation existed between the process-outcome assessments, thereby, in their estimation invalidated the process audit. Thus, it has become evident

that the choice among these process and outcome measures or the combined process-outcome assessments is still not clear. There is a lack of empirical evidence to differentiate the merits of these three measures. Quality of care refers conceptually to optimal performance by the medical system to produce the best possible outcome. Since it is difficult to determine what is optimal performance, quality of care will not be easy to measure no matter what method is employed (McAuliffe, 1979). Yet, as Brooks (1977) pointed out these contrasting views should not be used to call a moratorium on quality assessment activities until improved measuring tools have been devised. Instead efforts need to be concentrated on ameliorating the evident difficulties through further rigorous research in the field.

Measurement Theory

In past epidemiologic and evaluation studies of health care, clinical assessments have been used in the absence of more objective means to establish diagnosis or record events in the course of diseases. More recently, increased efforts have been made in the development of quantitative methods for the evaluation of health status, so that the assessment of a therapeutic response could be more objective.

As noted by Jette (1979), improved health is an undisputed universal goal of health practitioners. Evaluating the extent to which this goal is achieved continues to challenge care givers and researchers alike. In the health

measurement literature published during the last two decades, two fundamental questions continually re-emerge. What do we mean by the concept of health and how can we measure it? In response to these questions, a plethora of health status indicators for use in evaluating the effects of health services delivered to institutionalized and non-institutionalized patients with chronic disease have been created.

The concept that measurement is essential to scientific investigation is unquestionably accepted. The science of measurement has been primarily developed on the mathematical model. Over the years however, social scientists have extensively expanded measurement theory and adapted it to more abstract concepts in order to quantify subjective data (Blalock, 1968; Cronbach, 1951,1971; Nunnally, 1964, 1978). As a result, measurement can also be viewed as the process of linking abstract concepts to empirical indicants (Carmines and Zeller, 1979). The need to firm up "soft" data in the fields of medicine and health care research (Feinstein, 1980) has likewise led to the adoption of the measurement theory.

Measurement is thus described as taking a characteristic of someone or something, usually a behavior, attitude or attribute and putting it into a category or giving it a numerical value. The purpose of this process is to allow characteristics to be more precisely interpreted, compared, defined and manipulated. This can be achieved through the use of instrumentation which is a procedure of selecting or developing measuring devices or methods appropriate to a given

problem.

Frequently reference is made in the literature to the standardization of measures. Essentially, a measure is said to be "well standardized" if different people who employ the measure obtain similar results (Nunnally, 1978). For example, a measure of activities of daily living is well standardized if different health care workers who employ the methods, obtain similar numerical results for particular patients in specific occasions. Formulating explicit rules for the assignment of numbers is a major aspect of the standardization of a measure. When designing an instrument, the researcher must consider the conceptual focus to be achieved, the purpose or applicability, the quality of measurement and the operational approach of the available instruments (Dette, 1979).

Through scales of measurement, the variables to be studied can be clarified. There are four types of scales: nominal, ordinal, interval, and ratio. The simplest is the "nominal scale" which consists of two or more named categories or classes which are qualitatively different from each other. The next is the "ordinal scale", which ranks its categories along a continuum; thus each class bears the same situational relationship to the class which it follows. The "interval scale" also a rank ordering, is distinguished in that it possesses equal units of measurement, thus making it possible to interpret not only the order of scale scores but also the distance between them. The highest level of measurement is the "ratio scale" which has the properties of an interval

scale together with a fixed origin or point zero (Moser and Kalton, 1971). These scales are described in ascending order of power and preference. Each is stronger than the previous type, for it provides the kind of information furnished by the preceding type, but with certain additional information (Abramson, 1979).

In designing a measurement scale, specific criteria are required. The scale should be appropriate for use in the study, keeping in mind the conceptual definition of the variable and the objectives of the study. It should be practical, that is geared to the methods of data collection. It should be powerful enough to provide the appropriate details. The categories should be clearly defined, and sufficient in number. In addition, the scale should be collective exhaustive and mutually exclusive (Abramson, 1979).

Measurement has been examined in many ways. Each time, these questions are confronted. Is it reliable? In other words, is it an accurate, consistent, and stable measuring instrument? Is it valid? Put in another way is it really measuring what it is intended to measure and is it relevant?

It is not enough for a researcher to design an instrument based on common sense and logic. The principles of measurement theory must be followed to ensure that the test is reliable, that the precision of the test is acceptable, and that the test is valid or measuring what it claims to measure.

Reliability concerns the degree to which results are consistent or reproducible across repeated measurements (Carmines and Zellers, 1979). For example, an intelligence

test is reliable, if an individual obtains approximately the same score on repeated examinations. Any measuring instrument is relatively reliable if it is minimally affected by chance disturbances. Generally no two scores are exactly the same but it is the source and degree of variation that is important and the basis of the reliability theory.

In health care research it is essential that the reliability (reproducibility) of information be examined and possible sources of variation in measurements be identified. Several sources of variation can be present. These include changes in the characteristics being measured or lack of constancy; changes in the measurement instrument, that is the variation between instruments themselves or lack of precision, as well, as differences between the people collecting the information or lack of objectivity (Abramson, 1979).

Observer variation is a term which refers to variation arising from the persons making the observations, and not from changes in the characteristics being measured or from the measuring instruments. However, it is difficult to separate these concepts completely. The term is, therefore, used to indicate the differences between observations by different observers on different occasions (inter-observer variation) or by the same observers on different occasions (intra-observer variation). Although attempts must obviously be made to collect reliable data, it is important to remember that total reliability is neither possible nor essential (Abramson, 1979). As Abramson stated, it is unrealistic in studies of reliability to expect perfection. What is important, is to

know the degree and direction of the systematic variation or bias, particularly for those variables that play an important role in the investigation. By definition, the bias of an estimator is the difference between the average value of the estimates obtained in many repetitions of the study and the true value of what it is estimating (Anderson, 1980). Confounding factors or distrubing variables, (Abramson, 1977) are the major sources of bias. These background factors can seriously distort the estimate of the effect of independent or process variables in relation to outcome variables.

The precision of a measuring instrument is equally essential for assuring the accuracy of the measurement. Precision of a measure refers to the degree of change in the property under study that can be detected with a particular measurement procedure. Quantitative precision depends on a detailed specification of the phenomenon of interest (Jette 1984). Unless these factors are known, it may be difficult to avoid reaching unwarrented conclusions.

However, empirical measures that are reliable have only come half way towards achieving scientific acceptance. They must also be valid, that is, they must fulfil the purpose for which they are being used.

Validity, in constrast to reliability, is more of a theorectically oriented issue because it inevitably raises the question, "valid for what purpose?" Validity is thus defined as the extent to which any measuring instrument measures what it is intended to measure (Carmines and Zeller, 1979). For example, a driving test may be valid as an

indicator of how well an individual drives a motor vehicle but is invalid for other purposes, such as one's potential abilities in university. "One validates, not a test but an interpretation of data arising from a specified procedure" (Cronbach, 1971). This distinction is the core to validation because it is possible for a measuring instrument to be relatively valid for measuring one kind of phenomenon but not another. Therefore, we validate, not the measuring instrument itself, but the measuring instrument in relation to the purpose for which it is being used (Kerlinger and Pedhazur, 1973). Validity is usually a matter of degrees rather than an all or nothing property, and validation is an ongoing process. Consequently, new evidence may require modification of an existing measure or the development of a new and better approach to measure the attribute in question (Nunnally, 1978). Thus, when using indices which have been developed and validated in a particular setting, it is important to re-examine the validity of the instrument with respects to the variables that play an important role in the study at hand (Abramson, 1979).

There are four types of validity that must be considered. Face validity which concerns the extent to which the instrument "looks like it measures what it is intended to measure" is the lowest form of validity. Content validity the next type in ascending order of importance, concerns the plan and construct or the scope of the instrument. It asks the question, does it adequately sample all the elements of the composite variable it aims to measure? The third type,

criterion-related validity is based on findings that there is a correlation between the measure under consideration and another external measure. Finally the most important type, construct validity is evaluated by investigating what qualities a test measures and by determining the degree to which certain explanatory concepts or constructs account for performance on a test (Issac and Michael, 1981).

In sum, the utilization of the science of measurement in health care research permits an objective evaluation and comparison of test procedures and findings. It takes the guesswork out of scientific observation. Through numerical results from standardization measures, it is possible to report results in finer detail, minimizing personal judgements.

Indices

The most common conceptual focus in health evaluation indicators specifically designed for use with chronic conditions has been concerned with patient independence in activities necessary for daily living. Extensive studies have been conducted by researchers such as Carey and Posavac 1978), Inversen et al.,(1973), Jette (1978), Katz et al., 1972), Mahoney and Barthel (1965), Pfeffer et al.,(1982 , and Tourtellotte et al., (1965) to create assessment instruments. Each instrument examines different components of activity which can emcompass the physical, mental, social and/or functional capacity of individuals in a particular patient

population. However, only some of these instruments have been standardized in terms of validity and reliability for specific patient groups.

One of the original measures of activities of daily living (ADL) for use with chronic patients was developed by Mahoney and Barthel (1965). This is a numerical scale which examines the functional levels of independence in the physically disabled by assessing 10 factors related to self care activities and mobility. The Barthel Scale requires total independence for full credit in each category, weights each function separately and scores from 0, (total dependence) to 100, (total independence) for personal care. On each item the individual is given points for being able to perform an activity independently and fewer points for performance with help. The score values are weighted and may be 15, 10, 5 or 0. A score of 100 indicates the patient is able to provide personal care for himself in the home, although independent living may be limited by other factors. In the rehabilitation setting, the Barthel Index correlates well with clinical judgement and has been shown to predict both mortality (Wylie, 1967) and the ability to be discharged to less-restrictive settings (Granger and Greer, 1975).

In the past few years, ten factors of the Barthel Index have been expanded to cover more precise levels of activity (Granger et al., 1975, 1976, 1979) and it is currently called the Barthel Self-Care Ratings (Gresham et al., 1980). The factors are scored by assessing whether the patient can perform the activity independently, with help or supervision,

r not at all.

As seen in Figure 1-3, these categories demonstrate a predictable progression in the development of functional levels.

FIGURE 1-3

Barthel Index Scoring System

Barthel Score	Dependency Category
0-20	Total Dependent
21-40	Severly Dependent
41-60	Markedly Dependent
61-90	Moderately Dependent
91-99	Slightly Dependent
100	Independent in personal care but may not be able to live alone, perform housekeeping tasks, or meet the public

Numerous studies have employed this scale and have found that a score of less than 40 indicates that some help is necessary in the activities of mobility and self care. At a score of 60, most patients are independent in basic skills and half can dress, transfer and ambulate with assistance. Test retest correlations of 0.89 and an interrater reliability coefficient of 0.98 have been reported for the Barthel scale (Granger et al., 1979). In addition, Sherwood and her colleagues (1977) reported high alpha reliabilities ranging from .95 to 1.00 for three samples of hospital patients, suggesting that the test was consistent internally as a measure of self care activities. By definition, Cronbach's alpha, a synonym for internal consistency reliability is an estimate of the correlation between the total score across a series of items from a rating scale and the total score that would have been obtained had a comparable series of items been employed (Last, 1983). In terms of validity, the Bathel Index has been useful in predicting patients outcomes and correlates highly with other accepted daily living indices (Donaldson et al., 1973; Gresham et al., 1980).

Thus, this index has several advantages for use in a clinical trial. It is relatively complete in terms of ADL requirements and it is sensitive to detect small but real changes in function. In sum, the Barthel Index is probably the best known formalized functional assessment instrument in current American medical rehabilitation settings (Gresham and Labi, 1984).

Katz (1972), was instrumental in the creation of the Katz Index of ADL which is an ordinal scale measuring six outcome variables of daily living. Cases are ranked from A, (most independent) to G, (most dependent). This index was originally created to measure the effects of continued care for chronic patients in the community. It has a specific ranking of ADL functions which is in the reverse sequence in which functions are lost and regained in disease and senescence. This is of theoretical interest, but it does not compensate for the insensitivity to small but definite changes in groups of patients (Gresham et al., 1980). Although the rating is dichotomous, the way in which the observations are made permit a differentiation between those who are independent and those that can perform the activity but with help. Various studies have been reported which demonstrate that this scale is highly reproducible. Coefficients of reliability have ranged from 0.94 to 0.97, truly excellent consistency (Kane and Kane, 1981).

The Kenny Self-Care Evaluation (Inversen et al., 1973) is a numerical scale that scores seven major ADL categories, (bed activities, transfers, locomotion, dressing, personal hygiene, bowel and bladder and feeding). It is scored from 0 (dependent) to four (independent) in each major category and assumes each to be weighted equally. Therefore, it produces scores from 0 (totally dependent) to 24 which means a patient is independent in all categories. This scale, based on the assumption that nursing care requirements are the reciprocal

of functional deficits in ADL, avoids arbitrary weighting and is, therefore, less practical for post discharge monitoring (Gresham et al., 1980). However, the areas of reliability and validity of the scale were lacking in studies describing this index.

Pfeiffer's Multidimensional Functional Assessment Instrument, the OARS Older Americans Resources and Services Pfeiffer et al., 1976 specifically studies the problems and needs of the elderly patient. This index which is designed to assess individuals or groups of people extends beyond the physical problems and health service requirements. It includes sections to evaluate mental health, social, and economic problems as well as impairment in self-care capacity. It is scored from one, (out-standing function) to six (complete impairment) with each area being rated separately. Specific tests of reliability using the test-retest method for consistency and inter and intra-observer reliability have been reported and range from 0.34 for mental health to 0.84 for physical health in inter-rater agreement. Face and content were achieved through the construction of the index. Criterion-related validity using Kendall's Tau values ranged from 0.60 to 0.89 depending on the evaluators (Fillenbaum, 1981).

The LORS Index (Level of Rehabilitation) developed by Carey and Posavac (1978, and 1980) was designed to evaluate programs of physical medicine and rehabilitation and

specifically was used to determine whether discharged patients maintained the improvement made during their hospital stay. This scale which has been based on the Functional Life Scale (FLS, developed by Sarno et al., (197) has mainly been tested on stroke patients. It assesses orientation and memory, and also measures whether the patient can perform simple activities requiring cognitive function and both function assessed [illegible] indicates that the activity [illegible] indicates that the activities [illegible] Only those areas in which effects were most likely to bring out improvement through treatment were tested. An inter rater reliability of scale scores was equal to or greater than 0.96, with the higher correlations being seen between raters of the same professional discipline. Similarly, information obtained from either a nurse or a spouse demonstrated inter informant correlations of 0.82 for ADL and 0.81 for cognition. Cronbach's alpha was also used to examine the internal consistency which demonstrated mean homogeneities of 0.94 for ADL and 0.88 for cognition.

Jette's Functional Status Index (1978) has been used to examine the effectiveness of health care provided to the individual with polyarticular disability. This index defines function as being made up of related, yet distinct dimensions. They include the degree of help used, the degree of pain experienced, and the degree of difficulty in performing eighteen different activities of daily living. Performance is

scored on a scale of one, uses no help, to five, unable to do the activity. Scores for degrees of pain and difficulty range from zero, no pain or difficulty, to seven, severe pain or difficulty.

Earlier studies (Wylie, 1967; Granger and Greer, 1975; Anderson et al., 1976) carried the implicit assumption that an increase in dependence in the performance of an activity, constituted a loss of health. In contrast, the Jette scale attributes an increase in functional dependence to sometimes be a legitimate goal of health services provided to an uninstitutionalized adult with chronic disease like arthritis. For example, he argues that certain increases in dependence might be offset by other desirable outcomes such as a decrease in pain and/or level of difficulty in function. Therefore, he feels that the use of a cane to ambulate, ought to be considered within the context of the desired function.

In a second study, the Jette team (Deniston et al., 1980) used the scale in The Pilot Geriatric Arthritis Project to test the hypothesis that a multidisciplinary health team could improve the quality of life of older adults with arthritis. Comparisons were made with a well-known standard index, the Patient Classification Approach by Katz (1972) and Jones (1974). Both a concordance and intra-class correlation coefficient approach was used to assess the inter-observer reliability. Agreement ratio for dependence rating was found to be 75% for all but two items, however, ratios for degrees of pain and difficulty on ADL performance were generally lower. Most of the observer discordance could be attributed

to variability in interviewer interpretation of definitions and concepts of the index. The author concluded that further work on training and standardizing interviewers was indicated for further studies using this assessment instrument.

In general, the majority of these studies reported the development and management of new indices to measure functional status in patients. However, relatively few projects have approached the use of these tested instruments in a controlled clinical trial.

One exception is the Tourtellotte group (Tourtellotte et al., 1965; Kuzma et al., 1965, 1969; Petrie, 1975; Henderson, 1975). Over a period of 10 years, a battery of objective tests were developed to qualify certain neurological functions as cognition, strength, steadiness, reactions, speed, coordination, sensations, fatigue, gait, station, and selected skills of daily living. The long-term goal was to bring to clinical neurology, a type of quantification of the nervous system so that the results of therapeutic trials might be evaluated more objectively and hence be more valid. The tests have been extensively evaluated and used in several randomized double-blind trials in Multiple Sclerosis and Parkinsons disease. An example can be seen in the study conducted by Kuzma and Tourtellotte (1965), which investigated the reproducibility of a battery of neurological tests, the Quantitative Examination of Neurological Function. Using a "4 by 4" Graeco-Latin Square design with 2 observations per cell, data was collected to determine whether different observers

could obtain comparable results; whether the level of neurological function varied from day to day in repeated administration of the tests; and whether the level of neurological function varied during various periods of the day.

Two of the trained examiners were neurologists and two were physical therapists. Using analysis of variance for Graeco Latin Squares, the results indicated that different evaluators could be trained to obtain comparable results using the quantitative tests. Further, the level of neurological function scores obtained by these specific tests did not vary significantly during the four stated periods of one day and more importantly they did not differ significantly when the battery of tests were administered on the four consecutive days. The authors concluded that the battery of tests would provide a more objective means of assessing neurological function than the conventional examination, and thus, should have considerable merit when used in therapeutic trials.

Kuzma and colleagues (1969), followed with a study designed to assess the reliability of three instruments used in the evaluation of Multiple Sclerosis patients. The design of the experiment was an Incomplete Latin Square as described by Federer (1953) using five examiners and ten patients. Each patient was examined only three times at the beginning of the study and three more times six days later. Using analysis of variance for extended Incomplete Latin Squares, the results showed no significant difference among the evaluators on 82 of the 87 items used to measure neurological function. There was

no significant difference among the average values of the sequence of the 3 examinations nor among the average increments of change in the numerical scores between the first and second trials. Test-retest reliability coefficients for strength tests were the highest with coefficients in the 0.80 to 0.90 range. Speed and co-ordination coefficients were also high while those for two point discrimination were between the ranges of 0.50 and 0.60. Thus, the results of this study indicated that the evaluation methods were reliable in the examination of neurological status when used in clinical trials where several investigators contributed data.

Henderson (1975), was instrumental in describing an additional source of error. He was particularly concerned with training of all study raters when attempting to establish rater reliability. In this report, he discussed the possible variation of scores among program instructors as well as the study raters and suggested the importance of testing all examiners in a similar manner. The study conducted a consistency check of all raters engaged in their multicenter clinical trial. After extensive training, inconsistencies were significant in only four of the twenty-eight tests performed. Using test-retest methods, they claimed strong agreement for most of the tests which indicated little or no learning effects. In conclusion, the authors stressed that the training of all examiners responsible for clinical evaluations is critical to the successful use of indices in clinical trials.

Thus it is increasingly evident that the success of any health care research is dependent on assuring that the data collected are of good quality. Otherwise, it is unlikely that the conclusions drawn are reliable and meaningful. In the same manner, knowledge of the sources and extent of systematic variation for large scale clinical trials enhances the credibility of the study (Garraway, 1976; Jette, 1980; Kerner, 1981). Therefore, it is the belief that extensive and detailed quality control procedures must be developed in any clinical trial to assure good quality performance of all its participants. As Knatterud (1981), stated "concern for quality control procedures should begin in the early stages of planning and continue till the final study paper has been written. An error-free study is not a reasonable goal, but it is important that the number of errors is small and that errors incurred are randomly distributed among the study groups."

When conducting a therapeutic trial, there are many factors which influence the quality of data and thus the results obtained. Primary concern involves the measuring instrument and the manner in which the instrument is administered. Therefore, the issues of reliability and validity of the proposed measuring instrument must be considered. Yet, in examining the literature, one is struck by the relatively few articles addressing the measurement issues associated with the formation of reliable and valid health status indicators and the interrelationships among them. Most work has been directed towards the general

population (Brook et al., 1979; Chen, 1976; Patrick et al., 1973; Reynolds et al., 1974; Stewart et al., 1977; Wolinsky et al., 1980). Nevertheless, it is a fact that the measurement of health status among older persons is indeed, one of the most important research issues facing gerontology today, particularly when outcome variables are being used as indicators of health services utilization (Wolinsky et al., 1984).

Presenting at the NIH Technology Assessment Conference in 1983, Rubenstein pointed out that comprehensive assessment has become one of the cornerstones of geriatric medicine and suggested that improvement in diagnostic accuracy could lead to improvement in treatment. He continued by stating that it was a major objective of most geriatric assessment programs to avoid inappropriate use of services, especially those in institutional settings for reasons of compassion and cost. Reporting from the 1979 Report of the United States General Accounting Office, he stated that at least 10% to 20% of patients in skilled nursing facilities and 20% to 40% of patients in intermediate level care facilities received unnecessarily high levels of care. This was a wasteful use of scarce resources and this situation only created further disability by leading to premature labeling of a patient as irremedially ill. He concluded by stating that there was growing evidence that assessment could lead to improved appropriateness of placement.

Yet, apart from the development of a few frequently used indices of health status (Duke University Center for the Study

of Aging and Human Development, 1978; Katz et al., 1972; Lawton et al., 1969, 1982; Pfeiffer, 1975) and the self-rating scales of health status (Linn and Linn, 1982) gerontological studies, for the most part, have neglected the evaluation of these health status measurement issues among the elderly population. Thus, confusion remains as to the reliability and validity for the majority of measures used in determining the global and functional dimensions of health in the elderly and this must be remediated. On the other hand, measurement overkill is an ever present hazard. As Kane and Kane (1981), pointed out, there is a tendency in research endeavors to include a large number of scales as a substitute for careful targeting of ways to measure the desired outcomes of the program and this proclivity must be avoided.

The use of measurement scales in the assessment of the functional status of the elderly person has come to be essential both to good geriatric care and to investigations documenting the effect of various interventions. However, Kane and Kane (1981), provided a word of caution as to the dangers of over-interpretation or misinterpretation of outcomes. They stated that once a scale has been created to measure a complex and abstract quality, care must be taken not to reify the scores. While it is important to systematically assess important aspects of function, it is also essential that the instruments be chosen carefully based on a knowledge of the content of the instrument and the history of its use. Further, it is a fallacy to assume that instruments proven reliable and valid in certain specialized centers will

continue to be reliable in the hands of other researchers.

In summary, the health status of our elderly population and the consumption of health resources are two aspects of health-care management that demand immediate attention. Care-givers must continue to try and improve the health of this segment of the population and must do so in the most effective and efficient manner possible. Achievement of these goals can only be mastered through well designed clinical trials which focus on geriatric care and seek to obtain reliable and valid data.

CHAPTER II

The Parent Study

Description of the Facilities

The Royal Victoria Hospital (R.V.H.) is a 749 bed acute-care institution located in the core of Montreal. Affiliated with McGill University's Faculty of Medicine, it is an active teaching and research-oriented hospital. All major medical specialitites, other then pediatrics are offered. In addition, it is closely associated with the Montreal Neurological Institute and the Allan Memorial Institute. In the fiscal year 1979-1980, there were 20,175 adult admissions with an average length of stay (excluding psychiatric and chronic care patients) of 9.1 days per patient. In this period there were 44,046 emergency room visits.

The Royal Victoria Hospital has five medical floors. Of these, one floor is located in the private pavillion and is used mainly for elective and semi-urgent admissions. The remaining four general medical floors, have an average of 26.25 acute beds, and 6.75 self-care beds per floor. In addition, there are a number of speciality units: the Intensive Care Unit-Cardiac Care Unit (ICU-CCU), the Cardio-Pulmonary Investigatory Unit, Palliative Care, Dyalises, Dermatology, and Renal Transplant Unit. The self-care beds are used primarly for the elective admission of patients who do not require nursing care. However,

active-care patients can make use of these beds when direct nursing care is no longer required. Although the Tenth Medical floor does not have self-care facilities, it has access to the beds on other floors and since the occupancy rate for these beds averages around 40% to 50%, there is rarely a problem of access.

Each medical floor has a group of staff physicians in the Department of Medicine who in rotation are responsible for patient management. In addition, there is a medical house staff consisting of one senior resident, one junior resident and two interns. The staffing of nurses is service based. Nurses are assigned to specific floors according to their specialities. A social worker and a dietician are available for each floor while two physiotherapists and one occupational therapist service all four floors. Patients are evaluated by the professionals upon receipt of a consult from the house staff and each case is discussed at weekly social service rounds.

Emergency room admissions, either directly or through ICC-CCU, account for at least 90% of the active care bed-days on these floors. Because of the constant pressure from the emergency room, elective cases have little access to these beds. Direct admissions from the emergency room to the medical floors go in sequence. Thus, when beds are not available on a given floor, patients are temporarily transferred to an area with room until the designated floor is accessible. Exception to this procedure is seen in the example, when patients are transferred fron the ICC-CCU to the

medical floors. These patients account for 10% of the patient 70 years and over and can not go off-service. Therefore, they go to the first medical floor in the sequence when an empty bed becomes available. Rarely is a unit dropped from the sequential procedure, as a result of having many off-service patients. Nevertheless, if a floor does take a patient out of turn, it does not come up for an admission until the next round.

The Royal Victoria Hospital also takes in 50 to 80 social admissions a year through the emergency service. These are patients who have no apparent acute medical problems but who can not be cared for at home. Normally, these patients are located on the Eighth and Tenth Medical Wards, however, if the census is too high, these patients are transferred to other areas in the hospital so that the load is shared by all floors.

Description of Parent Study

As a result of a shortage of chronic care beds in the region of Montreal, patients in need of chronic institutional care are being served in the acute-care setting. According to government policy, the acute hospital with the greatest proportion of chronic care patients has first priority for transferring patients to specialized long-term care institutions. Because there are limited numbers of patients in this category at the Royal Victoria, this hospital rarely can make use of this opportunity to transfer patients, and

thus liberate beds for acute-care needs.

Therefore, in July 1980, a division of geriatrics within the Department of Medicine was opened at the RVH. This section differed from other divisions in the hospital because it was staffed with a multidisciplinary team which functioned as a consultation unit. Their duties were concentrated on the eight and tenth medical floors only, however consultations to patients on non-medical floors and out-patients were provided upon request. The aims of the team were to provide early assessment, treatment, and rehabilitation with the emphasis on improving functional and psychosocial status. Coordinated discharge planning was developed to provide services and care appropriate to the patients' and the families' needs. Furthermore, patient and family involvement was encouraged in the care process through individual counselling and family conferences.

Presented with the opportunity to examine the effectiveness of this new geriatric unit, a controlled clinical trial designed to study the effects of adding this geriatric consultation team to the traditional pattern of care for elderly patients was established on the acute medical wards. Patients over the age of 70 years who were admitted from the emergency room to two control floors (six and seven) and two trial floors (eight and ten) in the medical pavillion, were then followed for six months. The objectives were to determine if the geriatric team was able to effect favourable outcomes in the areas of: length of hospital stay; place of residence on discharge; physical, mental, and social

functional levels; and post-discharge consumption of medical services. Additionally, data on socio-demographic characteristics, functional status and diagnosis were collected to determine if certain patient characteristics were of value as predictors of outcome in this setting.

The admission criteria allowed inclusion of only residents of the greater Montreal area, eligible for Medicare benefits. Patients had to be 70 years and over and be admitted on an emergency basis directly to one of the four floors in the study. Furthermore, all patients had to sign or have a responsible person sign the informed consent form. Patients excluded from the study were those admitted electively, those in subspecialty beds as well as patients transferred from other floors, including ICU-CCU. These patients represented 10% of emergency admissions and for most of them, the process of care was dominated by a post-myocardial infarct proctocol. Patients admitted as social admissions to the two-trial floors were also excluded from the study. These individuals represented less than 5% of the emergency admissions to the trial floors.

Study participants were allocated from the emergency room to the four medical floors by traditionally established procedures as described earlier. Prior to the start of the geriatric team approach, preliminary data had been collected on all patients 64 years and over, discharged from these four study floors over a six month period. This information revealed no significant differences between the floors in total number of patients, mean age or mean length of stay.

Therefore, it was felt that the allocation of patients to the control and trial floors would be effectively distributed at random. To assess similarities or differences between the two groups, data was collected on socio-demographic and clinical characteristics, as well as physical and mental functional levels prior to and at admission.

There were two clinical plans provided to the patients selected for the study. The program for the Team Care approach, as mentioned previously, was a multi-disciplinary effort aimed at improving the process of care for the elderly patient. The Geriatric Team consisted of two Internal Medicine physicians, one nurse consultant, one physiotherapist and one occupational therapist, one activity counsellor, and the social worker assigned to the specific floor. The initial evaluations were conducted by the physician and resident consultants to identify problems common to this population such as falls, incontinence, decreased mobility, dementia, confusion, and pressure sores. In addition, patients demonstrating functional disabilities were referred to the team unit through the nurse consultant and/or the physical or the occupational therapist. Emphasis was placed on early assessments and treatment of the patients needs including the functional status and psychosocial levels. During weekly multidisciplinary conferences, all new patients were discussed and the progress of current patients reviewed. Members of the Team followed the patients regularly to ensure that the selected programs of therapy were implemented. In addition, attempts were made to organize discharge plans to meet the

needs of the patients and their families. In doing so, liasons were created with other institutions and community services to provide follow-up care for their clientele.

The program for the Traditional Care approach for the control group of patients consisted of care as it conventionally had been given on the wards at the Royal Victoria Hospital. The Primary Care physicians were responsible for the care and the plan of treatment for each of their patients. Specialty consultations and services such as social services, occupational or physical therapy were only provided by each discipline upon request from the treating physician. As a result, these professionals generally worked independently of one another, communicating primarily through the patient's dossier and weekly social service rounds.

The study admitted 404 patients 182 to the control floors and 222 to the trial floors. Patients entered the study through a screening process conducted by the research assistant shortly after admission to the medical ward. Two to four days follow admission pre-test measures of some of the dependent or outcome variables were made. Self care skills, mobility, and mental status were assessed by one of the 14 trained evaluators. These evaluators were selected from outside the Royal Victoria Hospital and were unaware of the study objectives or the treatment group of the patients.

The collection of both process and outcome data was considered necessary to assess the quality of care delivered and the effectiveness of the program offered. In terms of the process of care, information was gathered on the time spent in

the hospital for both groups of patients in an effort to calculate the geriatric team's involvement. Special nursing procedures, complications, and consultations were recorded as well as the timing of implementation of specialized services of physical and occupational therapy. The use of health services after discharge was monitored on an ongoing basis with specific information being collected on the utilization of physician and community nurse services, social service contacts, hours of home services, and visits to the emergency.

Four follow-up interview evaluations were conducted by the same trained evaluators who completed the initial assessments either by telephone or direct contact, wherever the person was at the time. These were at 12 to 16 days post admission, one month, three months, and six months post admissions. The outcome variables provided information on the patient's ability for self care, continence, mobility. In addition, home activities, social interaction, and mental status of the individual were also monitored.

The instruments used in this study included the Barthel Index (Mahoney and Barthel, 1965; Granger et al., 1979) to measure performance in activities of daily living (ADL); the LOPS (Level of Rehabilitation Scale, Carey and Posavac, 1978) to measure activities and social interaction and the Portable Mental Status (Pfeiffer, 1976), a subscale of the OARS "Multidimensional Functional Assessment" to measure mental function. All instruments selected for this study had previously been field tested and standardized on a variety of patient populations. Therefore, it was felt that they would

provide a good mechanism for observing and reporting the patients' functional abilities and changes over time. In addition, data on behavior patterns such as wandering, aggression, and abusiveness were collected. Although this index was not standardized, knowledge of behavior patterns was considered important because of the impact it had on place of residence upon discharge.

In the spring of 1984, the final patient evaluations were completed. The data collected was coded and processed for analysis. Through this breakdown, it was possible to determine whether the two study groups were, in fact, alike with respects to their baseline characteristics. The two groups were then examined and compared for their outcomes from hospital stay. In addition, specific analysis was used to define which baseline factors, other than team care, were associated with favourable or unfavourable outcomes. This information will aid in the development of a system which can be used to recognize future high and low risk patients.

The acute care hospital plays and will continue to play a predominant role in meeting the needs of the elderly patient. Therefore, it is important to clearly define this role and the effectiveness of the services offered to this population. Thus, the study conducted at the Royal Victoria Hospital was able to provide vital information on issues of individual patient management, and guidelines for the future planning of services.

CHAPTER III

The Present Study

The present study examines the quality of the data collected from the Parent Investigation. The over-all objective of this research was to assess the reliability and the validity of a set of standardized instruments which were used in the controlled clinical trial of elderly patients. Given the distinct elements of both sets of measurements, two separate studies were conducted and therefore will be individually reported.

Quality Assessment

The Reliability Study

Objectives and Hypothesis

The specific objectives of this reliability investigation were threefold: a) to estimate the overall variation among the twenty-three raters and the study norm, in the scoring of selected functional status indices. b) to examine the inter-observer reliability among the fourteen evaluators, five interpreters, and the four study instructors. c) and finally, to measure the intra-observer variation for all raters in the two trial investigation.

The initial experiment examined the degree of the over-all variation in the observed scores among the study evaluators, interpreters, instructors when compared to the norms of the reliability study (Dr. S.W-Dauphinee and Mme M. de Lormier). The Prinicpal Null Hypotheses were: a) There was no significant difference in the mean scores among the evaluators, interpreters, instructors, and the norm reference. b) There was no significant difference between the study raters. That is, on the average, the examiners obtained uniform scores on the same patients assuming that the patient status did not alter. c) There was no significant difference within the same rater as measured in the test-retest sequence (second to first trial).

In recognition of possible systematic bias among study educators (Henderson, 1975; Knatterud, 1981) the fourth hypothesis was included which stated: d) There was no significant difference between the mean scores of the four study instructors; Dr. D.G., Principle Investigator of the Geriatric Study; Dr. S.W-D., Co-Investigator; Mme M. de L., Study Co-ordinator; and Mme S.J.B., Study Instructor.

Study Populations and Methods

Selection of Participants

The study raters were classified into three separate groups. The first group was referred to as the Parent Study evaluators and were responsible for the collecting of data for the Geriatric Trial. The interpreters made up the second group of raters and were employed by the Parent Study to assist the reseach assistant and study evaluators in communication and data collection procedures with patients who did not understand English or French. The third group of raters consisted of the study instructors who were responsible for the operation of the Geriatric Study and the education of the selected raters.

Evaluators

Since it was considered important that the study raters have an understanding of, and experience in the art of interviewing, study evaluators were recruited with these attributes when possible. Prior knowledge and practice in the measurement of functional status in a patient population was also preferred. This request resulted in the employment of several candidates with professional backgrounds in the fields of nursing, physical and occupational therapy. Careful attention was given to the selection of data collectors not associated with the Royal Victoria Hospital to reduce the

awareness of the study objectives. Nevertheless, two evaluators were selected from this institution, and therefore were responsable for gathering data by telephone or home interviews exclusively.

In total sixteen evaluators were chosen for the program. There were five nurses, four physical therapists, two occupational therapists, and five non-professional individuals. Of these, two people later refused to take part in the initial reliability experiment and therefore were dropped as data collectors from the Clinical Trial leaving 14 active evaluators.

Interpreters

Because of the heterogeneous catchment area of the Royal Victoria Hospital, patients with different ethnicities were frequently admitted. To deal with this factor several interpreters were recuited to assist the research assistant in collecting pertinent data. In addition, these interpreters provided an essential link between the study evaluators and many admitted patients. Through simultaneous translations of the study questionnaires, many patients who had accepted to become a part of the Geriatric Trial, continued to participate in the program over the required six month follow-up. The lanuages needed for this study included Chinese, Portuguese, Italian, Greek, Polish, French, and English. Thus, five different interpreters were selected, trained, and made available for all the necessary hospital, home, and telephone

interviews.

Patients

In order to conduct the reliability experiment, six patients were selected from the "Centre de Réadaptation de Constance Lethbridge" to take part in a series of video-taped interviews. The criterion for selection required that the subject have some degree of physical disability but good mental function. After each patient was given careful explanation of the purposes and procedures of the video sessions, informed consent was obtained from each participating individual (Appendix 1). All patients were advised that they could withdraw from the testing program if they so desired and were assured that their departure would not be detrimental to their continued care.

Training the Raters

Careful training of raters is critical to the successful use of functional status indices in any large clinical trial (Henderson, 1975). Therefore, prior to the start of the main study, a detailed training regime was instituted at the Royal Victoria Hospital to familiarize all study raters with the study population, the selected study indices, and their scoring systems (Figure 3-1). Since not all evaluators were experienced in patient illnesses and hospital procedures, an introductory session reviewing the more common geriatric

illness was provided by Dr. D.G. one of the study instructors. The trainees were then given complete written descriptions of the scales to be used, operating procedures, and specific instructions in how to conduct an interview. On the same day, one of the trainers (Dr. S. W-D or Mme M. de L.) demonstrated each scale with selected consenting patients and independent scores were recored by all raters. The examiners were then instructed to study their work manuals and to return for further testing at a later date. During the second training session, each examiner was requested to conduct interviews with two or three different patients while the trainer observed. Both trainer and trainee independently recorded the scores for each of the testing scales and findings were later examined for their consistency. Results within a ten point spread were required, the exact agreement depending on the specific test. Major discrepancies were discussed with each evaluator to determine their interpretation of the particular question and to gain further insight into the evaluator's over-all understanding of the index.

FIGURE 3-1

TRAINING REGIME

Introductory Session

Training Session

Use of Scales Interviewing Techniques

Patient - Trainer Interviews

Patient - Trainee Interviews

Video Testing Session I

Video Testing Session II

Reliability Study

Overview

Original plans for this present research were to encorporate an Incomplete Latin Square design in the initial evaluation of the reproducibility of the study raters as reported by Kuzma (1969). However, with the expansion of the number of data collectors from eight to nineteen, and the severity of the illness of the study population under consideration, the Incomplete Latin Square design posed several problems. In addition, reported patient fatigue during testing interviews (Kuzma et al., 1969), learning effects on the part of patients and raters (Potvin et al., 1975; Henderson et al., 1975; Loewenson et al., 1972), and administrative and scheduling problems confirmed the need for a design which would control for these sources of error and not tax the target patient population unnecessarily. Therefore, a factorial design was selected for this experiment through the aid of the video-taped patient interviews. By definition, the factorial design is a method of setting up an experiment to assure that all levels of each intervention or classificatory level is used in combination and sequence with every level of the other study factors (Last, 1983).

Study Description

As mentioned, the reproducibility experiment encorporated a series of video-taped interviews consisting of six patients selected for the variability of their physical status. The "Centre de Réadaptation de Constance Lethbridge" was approached to request their assistance in the preparation of these audio-visual tapings. This Center specializes in the long-term rehabilitation of patients suffering from varied physical and mental disabilities. Emphasis is placed on regaining degrees of functional independence and the return to community living. It was felt that the climate at this Center was more in keeping with the home environment and therefore, could provide an appropriate setting for the testing of the study scales, particularly when examining the variables of ambulation, home and outside activities.

Six consenting patients attending the Center were selected for the taping sessions: three French-speaking and three English-speaking individuals. In order to insure complete understanding between the interviewers and patients, two separate trainers or instructors were used in the taping sessions. Dr. S. W-D. was responsible for the English interviews while Mme M. de L. conducted the French sessions. From these original tapes, standard scores were recorded to be later used in the analysis of the over-all variation between the study raters and the pre-determined study norm.

A series of tests were then organized using the video-taped interviews to evaluate the consistency of scores

for all study participants. Each evaluator and interpreter viewed all six patient tapes, three French and three English and independently scored the patients with the appropriate indices. Since direct patient interviewing was eliminated from these sessions, questions concerning points of ambiguity or lack of clarity were answered by the study instructor responsible for the preparation of the video tapes. The order of tape presentation followed the one through six sequence. Two weeks later, the tests were repeated as in the first session but the order of the tapes was reversed. Each observer rated the six randomly selected patients for each of the selected order conditions, so that a total of 168 different test performances were observed for the fourteen evaluators and 60 test performances were reviewed for the five interpreters. Henderson, (1975) identified further possible measurement errors steming from inconsistencies among study instructors themselves, therefore putting the entire educational program and training sessions in question. With this in mind, the four study instructors were also required to view all six video-tapes, scoring each patient independently as stipulated in the testing sessions for all other evaluators. Each instructor rated the six randomly selected patients for each of the selected order conditions, so that a total of 48 different test performances were obtained for the instructors. The objective of this exercise was to determine if agreement could be attained among the study trainers who were responsible for the continued instruction of the group of trainees. If a major deviation of scores was seen in either

trainers and trainees, a second training session would be required and a second reliability trial set-up.

Assuring Adherence to Study Protocol

The monitoring of adherence to study protocol was assured through one hundred and five inspections of hospital interviews and forty-six inspections of home evaluations. During these inspections, an instructor accompanied the evaluator and the interpreter if present and independent scores were recorded by all raters. Hospital interviews averaged between ten and twenty minutes in length, while home interviews generally took twenty to thirty minutes to complete. This additional observation period provided an on-going verification of the consistency or variation of scores and on-the-spot instruction, if difficulties were encountered.

Instrumentation

1. Barthel Index

In the Geriatric Study, the performance of activities of daily activities (ADL) was measured by the use of the Barthel Index, developed by Mahoney and Barthel (1965) (Appendix E). Through this scale, a baseline for each patient was established, and progress in a treatment program followed. This scale likewise, served as one of the testing tools for the reliability experiment.

the reliability experiment.

2. Level of Rehabilitation Scale (LORS)

To examine the "quality of life" of the individual in the community, the Geriatric Study selected the Level of Rehabilitation Scale (LORS) (Carey and Posavoc, 1978) (Appendix 3). This scale evaluated the basic household functions related to living at home, as well as engaging in outside activities and social interaction with other individuals. However, only seventeen of the eighteen items were used as it was felt that the question concerning work and school activities was generally not applicable to their elderly population.

Data Collection Procedures

Data for the Geriatric Trial were obtained through hospital, home or telephone interviews with patients and a relative when possible or another 'significant person' using the two study scales. Similarly, these scales were used in the Reliability Study to examine the variation among the data collectors, and to determine their general understanding of daily living activities. Data on pertinent characteristics were obtained for all raters participating in the Reliability Experiment. This was achieved by assigning one of the study instructors, the responsibility of scheduling and testing each of the selected raters. Data were gathered, verified, and coded for each rater during the video-testing sessions and

during the periodic inspections to ensure adherence to study protocol. Repeated examinations of each rater were possible through scheduled patient interviews either in the hospital or home setting.

Detailed records were kept throughout the study on each rater as well as those who dropped out or refused to continue in the testing procedure. All data were coded twice and randomly checked a third time to ensure accuracy in the recording of the information. The codes were then copied to transcription sheets and verified for copy error. Professional computer data entry services were used to reduce further possibility of mistakes in filing the data into separate working files in preparation for analysis.

The Validity Study

The specific objective of this aspect of the investigation was to determine the validity of set of selected study indices chosen by the Geriatric Trial of the Royal Victoria Hospital to evaluate the functional status of persons aged 70 years and over.

The Principle Null Hypothesis was: There was no significant relationship between the scores from the Geriatric Study Indices, (Barthel Index by Mahoney and Barthel, (1965) and the (LORS) Level of Rehabilitation Scale by Carey and Posavac, (1978)) and the selected validity testing scale (Functional Status Assessment Instrument (FSAI) by Jette, (1978)).

Validity Design

This present study examined the validity of the LORS and the Barthel Index in association with the Functional Status Assessment Instrument (FSAI) designed by Jette, (1978) using criterion-related validation techniques (concurrent-type).

Program Description

The Validity Study was divided into two parts. The first section dealt with the Hospital Interviews only which included patients who were initially evaluated in the acute-care setting and continued to remain in hospital or some other form of semi-protected environment for the full six month follow-up of the Geriatric Trial. For these interviews, the Barthel Index and the mobility and self-care items of the Functional Status Assessment Instrument (FSAI) were compared. The LORS was not included at this time, as the tasks of food preparation, home maintenance and outside activities were generally not applicable for these people.

The second section of the Validity Study examined patients initially seen in the acute-care setting and then followed through personal interviews in their own homes. For this aspect of the study, the Barthel Index was used to collect data on mobility and self-care while the LORS was used to record home, outside and social activities. In this instance, the full Functional Status Assessment Instrument (FSAI) was employed to test the validity of the Geriatric scales using criterion-related procedures. In order to limit the problem of patient fatigue during the scheduled hospital and home interviews, the evaluator conducted the interview using the two Geriatric Scales while the accompanying instructor simultaneously recorded the responses on all three scales. The FSAI included some questions which were not present on either the Barthel Index or the LORS. These

questions were then asked after the formal interview had been completed. It was felt that the additional information from the instrument (FSAI) would provide further detail of the patient's functional status.

Instrumentation

Functional Status Assessment Instrument (FSAI)

The Prinicple Instrument of the Validity Study was the Functional Status Assessment Instrument (FSAI) designed by Jette (1978). The current version of the FSAI assesses dependence, pain experienced, and difficulty involved in performing eighteen different activities of daily living. These 18 activities were selected from an original pool of over 45 activities based on factor analyses of the underlying structure of these items (Jette, 1980). The modified version was then tested for its reliability in a study involving adults with rheumatoid arthritis (Jette, 1980). In responding to questions in the FSAI, a respondent is asked to use the time frame of the previous seven day period. Respondents are further asked to respond according to their preceptions of the average amount of help used, pain experienced, or difficulty involved in carring out their daily routine. Thus the FSAI is designed to gather a relatively stable assessment of average function over a relatively short time period. The scoring system for dependence ranges from one (1) 'uses no help' to five (5) 'unable or unsafe to do the activity'. Pain is

measured on a scale of zero (0) 'no pain' to seven (7) 'extremely severe pain', and difficulty on a scale of zero (0) 'not difficult' to seven (7) 'extremely difficult' (Appendix 4).

Data Collection Procedures

Similar techniques were used in the handling, coding, and verifying of the Validity data as was employed in the Reliability Study.

Data Analysis

The data recorded on the three study questionnaires (Appendices II, III, IV) were then registered on OMR Data Coding Sheets, NCS Trans Optic MB08-32768-7654 and subsequently entered into the McGill Amdahl 5850 Computer. The data were verified by visual inspection with the original questionnaires as well as by machine implemented frequency and logical checks. All detected errors were corrected and the cleaned data were then set into separate files in preparation for analysis.

Reliability Study

The primary hypothesis of the Reliability Study, that there were no differences between the fourteen evaluators, the five interpreters, the four instructors and the study criterion was tested using the analysis of variance. Similarily, the three remaining hypotheses, that there were no differences between and within the twenty-three study raters was tested by the analysis of variance, using the mixed model for nested classifications with unequal samples. Analysis of variance (ANOVA) is defined as a statistical procedure that isolates and assesses the contribution of categorical factors to variation in the mean of a continuous outcome variable. The data are divided into categories based on their values for each of the independent variables, and the differences between the mean outcome values of these categories are tested for

statistical significance (Last, 1983). A mixed model (Model III) applies when the experiment involves one or more factors which have fixed levels and the remaining factors are a random sample of a set of treatments about which the experiment wants to make inferences (Hays, 1963). Nested by definition occurs in an experiment when categories or levels of one factor take effect only within levels of another factor (Hays, 1963).

The overall variation between raters is provided by the analysis variance table and its "F test". However, to examine the differences between the groups of participants a more detailed breakdown of the independent variables is required. Instead of estimating effects directly by taking differences of the treatment means from the grand mean, as in fixed effects models, in mixed models the interest lies in estimating the true variance attributed to the examiners. The proper tool for determining appropriate error variances for more complex situations is the set of expected mean squares. Expected mean squares are algebraic expressions specifying what functions of the model parameters are estimated by the mean squares resulting from partitioning the sum of squares. Generally, the expected mean squares are linear functions of elements representing: (1) error variance, (2) functions of variances of random effects, and (3) functions of sums of squares and products (quadratic forms) of fixed effects (Freund and Littell, 1981). Using the SAS Program for General Linear Models, the coefficients of the components of variance and the subsequent equations for the expected mean squares can be obtained from the Analysis of Variance. Through the

partitioning of the components of variation for selected dependent variables, the differences between and within the study raters as well as the observer-order interaction can be determined.

One way of demonstrating that a factor accounts for a given amount of variance is by the index known as the population intraclass correlation coefficient (Bartko, 1966; Fleiss, 1975; Kramer and Feinstein, 1981):

$$\rho_I = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_E}$$

The intraclass coefficient for the grand population will be zero when σ^2_A is zero, and will reach unity only when $\sigma^2_E = 0$, given that $\sigma^2_J \neq 0$. This intraclass correlation coefficient has been used in this study to determine the inter and intra-rater reliability of all study observers. In addition, the preliminary data analysis included mean ±SD, percentage of agreement ratios, meaurement bias, and Pearson correlation coefficient to examine the trends between associated variables in comparison to the concordance between these same variables.

Validity Study

For the validity study, Spearman Rank Correlation Coefficients (Rho) were employed to examine the degree of association between the three test scales. The Spearman Rank Correlation Coefficient is a nomparametric measure that is calculated as the association of the ranks of the data (Sall

and Delong, 1982). This correlation indicates the degree to which two or more sets of observations fit a linear relationship. This coefficient represented by the letter "r" can vary between +1 and -1. If r=+1, there is a perfect linear relationship in which one variable varies directly with the other. If r=-1, there is a perfect linear association but one variable varies inversely with the other (Last, 1983). The Barthel Index and the LORS both are scales which use nominal weighted classification for the separate items of function but provide a total score in continuous terms. Whereas, the FSAI is a continuous scale ranging from one through five for dependence and one through seven for pain and difficulty. Due to the differences in the scoring systems on these three scales, nonparametric tests were chosen as few assumptions are made about the properties of the parent distrubtions.

All procedures performed for these studies employed the Statistical Analysis Systems (SAS). The SAS GLM procedures for the Analysis of Variance (Goodnight, 1982) and the SAS Correlation Procedures for Pearson and Spearman's Correlation Coefficients (Sall and Delong, 1982) were used.

CHAPTER IV

Results

Part I: Reliability Study

In total 288 video observations were examined to determine the reliability among the twenty-three study raters when compared to the Gold Standard. These video observations were made by three distinct groups of raters: Group I was comprised of 14 evaluators, Group II had five interpreters, and Group III consisted of four instructors. Using the Barthel Index and Level of Rehabilitation Scale, data were collected and compiled for the functionally related areas of mobility, continence, self care, home activities, outside activities, and social interaction. Scores for each group of raters were compared by a series of statistical procedures.

Comparison of the Three Groups of Raters to the Gold Standard

The scores obtained by the three groups of raters were compared to the Gold Standard for the two video sessions (Figure 4-1, 4-2). The distribution of the scores for the six patients demonstrated that differences in overall physical functioning were apparent for each of the patients selected. In each case, the Gold Standard expressed by the dotted line, set the measure of performance for each of the six subjects.

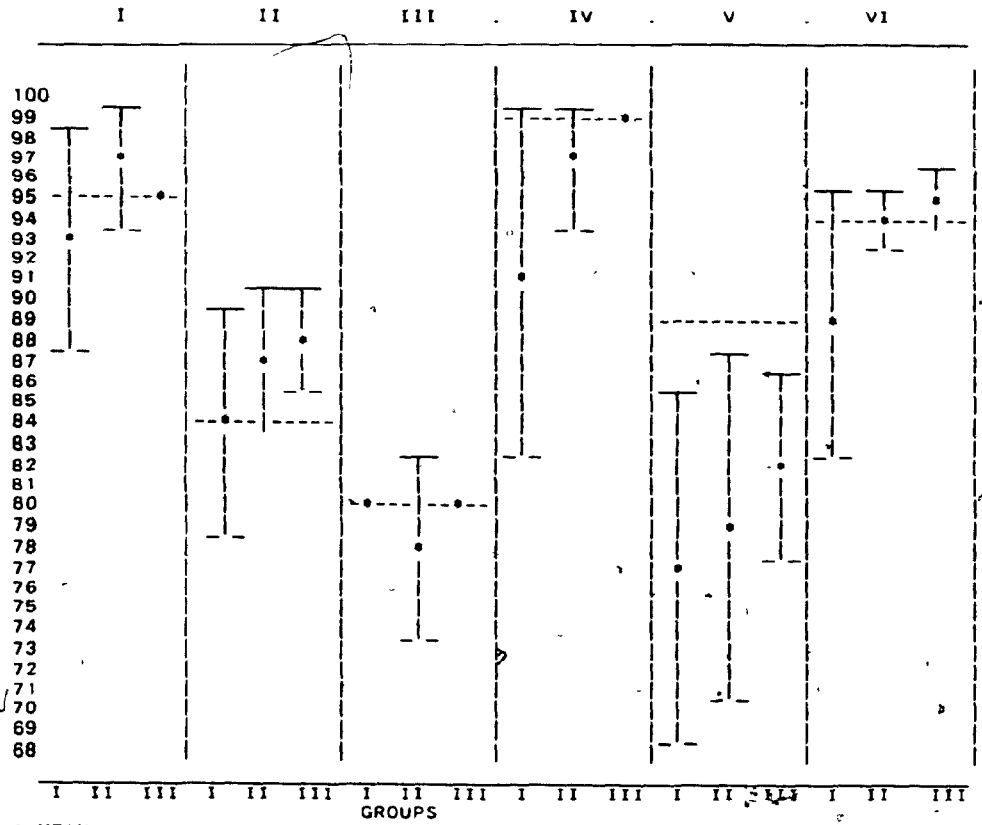
Through the plotting of the means and the Standard deviations of all the observed values, it became evident that variability was present among the three groups of raters when compared to the study's gold Standard. The question was how much?

In the first video session, the distribution of the group means was moderately consistent for the first three subjects while a greater variation among the raters and the Standard was present for patients, "four through six". On average, the greatest variation of scores was seen among the Evaluators (Group I) while the Instructors (Group III) were credited with the least deviation. Interestingly, the scores recorded by the Interpreters (Group II) tended to fall between the two Professional Groups. When examining the repeated scores of the three groups of raters with the Standard, a corresponding pattern of results was obtained. The Evaluator scores, although demonstrating the greatest range of variation remained relative consistent over the two testing sessions. The Interpreters, on the other hand, and to lesser extent the Instructors showed a slight increase in the spread of functional scores. Here again, the greatest indecisions among the raters involved patients, "five and six". In general, there was a trend towards an underestimation of patient status and this was more prominent among the Evaluators, however, all raters underscored patient number "five".

MEAN SCORES AND STANDARD DEVIATIONS FOR THE THREE GROUPS

OF RATERS USING THE BARTHEL INDEX

PATIENT VIDEO TAPES SESSION I



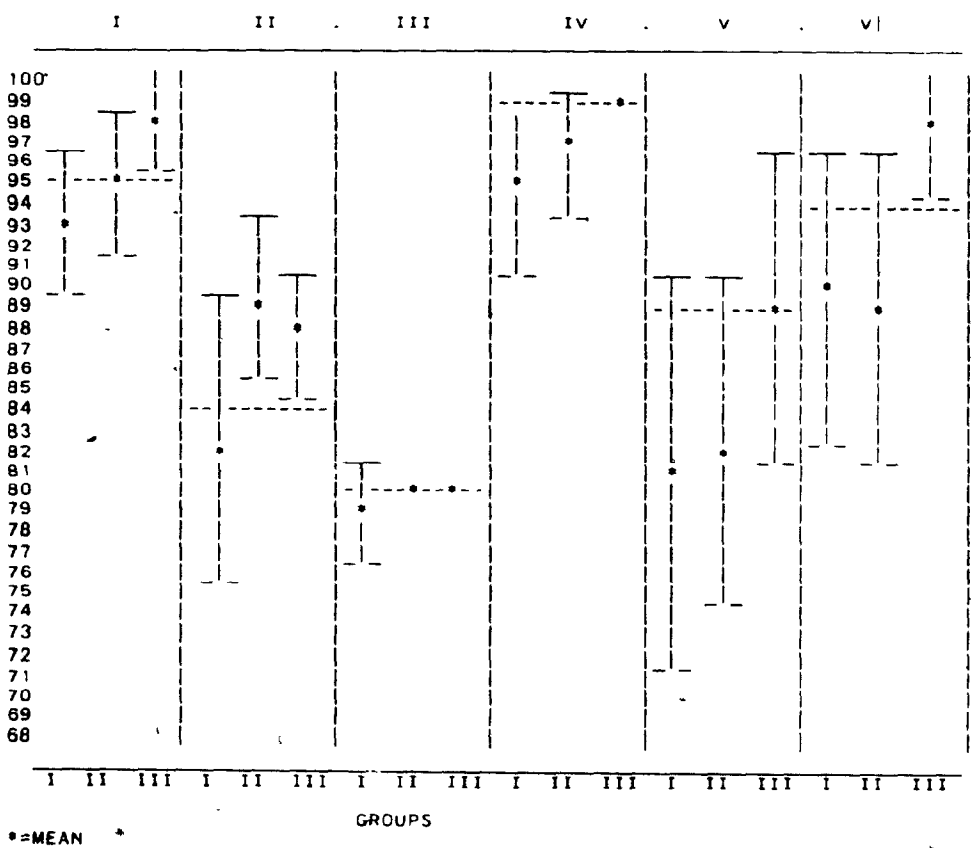
•=MEAN

DOTTED LINE ___=GOLD STANDARD SCORE

Figure 4-1

MEAN SCORES AND STANDARD DEVIATIONS FOR THE THREE GROUPS

OF RATERS USING THE BARTHEL INDEX



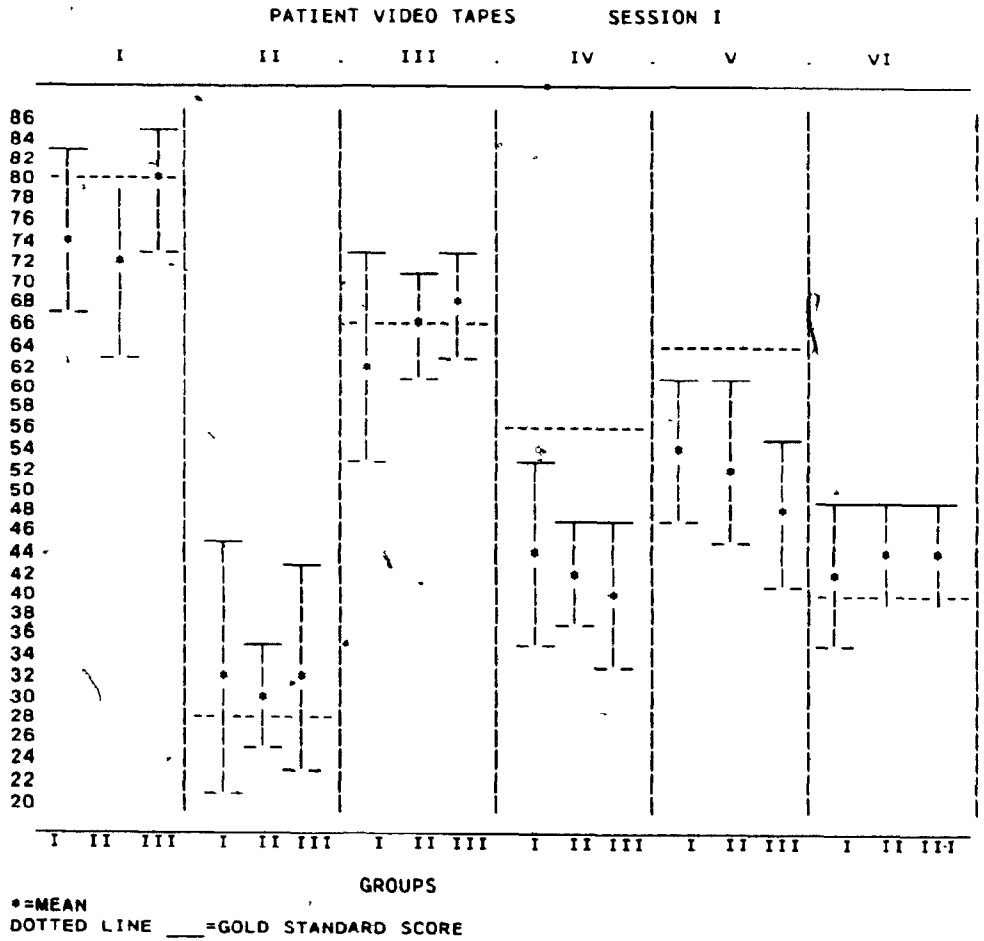
GROUP I EVALUATORS
GROUP II INTERPRETERS
GROUP III INSTRUCTORS

Figure 4-2

In a similar fashion, using the Level of Rehabilitation Scale to measure the instrumental activities of daily living, the means and Standard deviations were compared between the three groups of raters and the Gold Standard over two video testing sessions (Figure 4-3, 4-4). The total score is expressed as a percentage, 100% representing independent function. The Gold Standard was again represented by the dotted line, setting the level of function for each of the six patients. Upon consideration of these figures, several findings were apparent. First, there appeared to be a greater concensus of opinion among the three groups of raters and the Gold Standard for three of the six participating subjects. While there was a spread of 10 to 14 points separating the raters and the Standard when patients "one, four, and five" were interviewed; it was patients "four and five" who were seen to be the principle sources of rater variability.

In comparing plots of the two video testing sessions, a decrease in the overall rater variation with the Gold Standard was noted. Underestimation was again apparent for three of the six patients tested however, fairly good agreement among the raters was evident for the remaining three subjects. In addition to these findings, there appeared to be a tighter clustering of group means for all subjects evaluated. In constrast to the Barthel Index, the LORS seemed to promote a better agreement between the raters and the Gold Standard.

MEAN SCORES AND STANDARD DEVIATIONS FOR THE THREE GROUPS
OF RATERS USING THE LORS INSTRUMENT

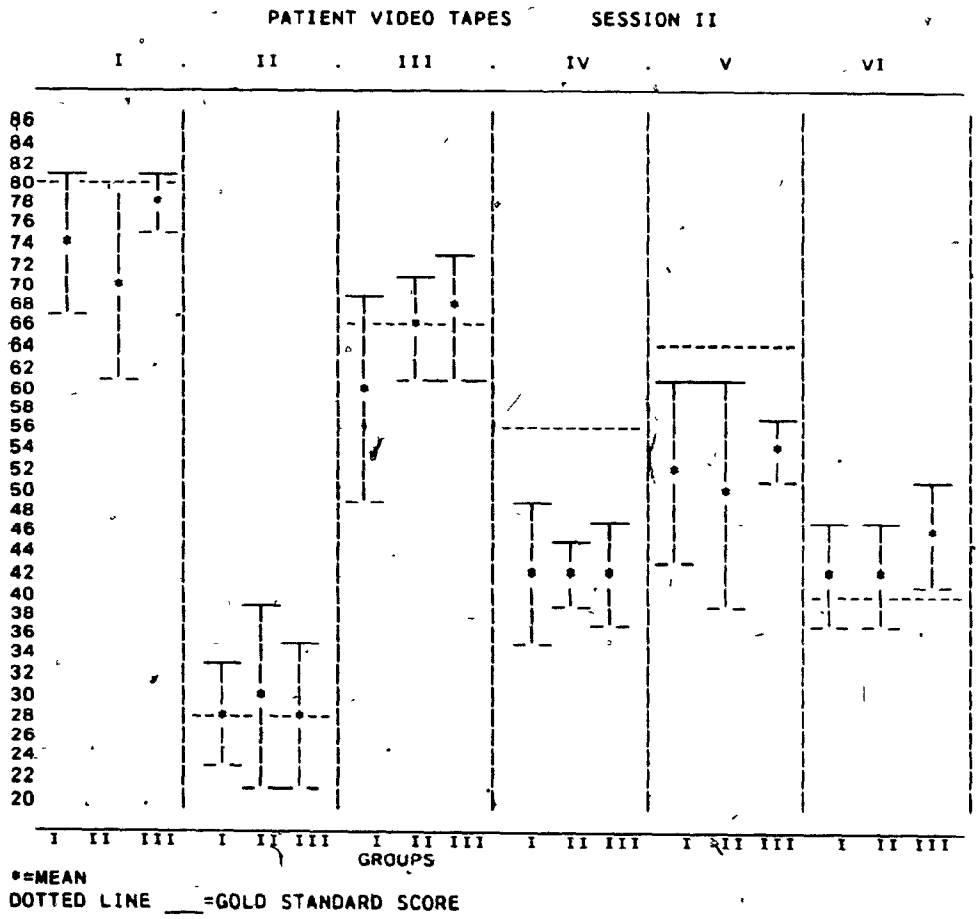


GROUP I EVALUATORS
GROUP II INTERPRETERS
GROUP III INSTRUCTORS

Figure 4-3

MEAN SCORES AND STANDARD DEVIATIONS FOR THE THREE GROUPS

OF RATERS USING THE LORS INSTRUMENT



GROUP I EVALUATORS
GROUP II INTERPRETERS
GROUP III INSTRUCTORS

Figure 4-4

•=MEAN

DOTTED LINE ___=GOLD STANDARD SCORE

Percentage of Agreement

To clarify the raters understanding of the study scales, both indices were broken down into their component subscales. The Barthel Scale was divided into three distinct subdivisions representing personal self care, continence, and mobility, whereas, the LORS was separated into household activities, outside activities, and social interaction. To identify the specific sources of variation among the three groups of raters and the Gold Standard for the six study patients, detailed analyses of each subgroup were performed. The first procedure used was the concordance approach. The expression of percentage agreement or agreement ratio has been the traditional way of indexing concordance for ordinal data (Kramer and Feinstein, 1981). For this study, the principle interest was to determine the percentage of agreement between the twenty-three raters and the Gold Standard. The agreement ratio is defined as the number of raters in accordance with the Gold Standard per item, divided by the total number of observations for that item.

$$\text{Agreement Ratio} = \frac{\#\text{Raters agree with the Gold Standard}}{\text{total \# of observations}} \times 100\%$$

Using this formula, it was possible to compare the Barthel Index data as scored by the three groups of raters and the Gold Standard over the two testing sessions. Findings are reported in Table 4-1 for the three subscales and the Total

score of the Barthel Index.

The first variable to be assessed was that of Personal Self Care. In the first testing session, the percentage of agreement between the twenty-three raters and the Standard ranged from 50% for patient "five" to complete agreement 100% in the case of patient "one" and patient "three". As seen in the presentation of the mean scores, patient "five" continued to produce the greatest source of uncertainity among the raters. It is not surprising, therefore, that the agreement ratio between the raters and the Standard was relatively low for this patient.

The Continence variable was then examined. In this situation, the calculated agreement ratios appeared to be good to excellent with values ranging from 83% to 100%. Once more, it was the performance of the fifth patient which reduced the level of overall agreement.

The level of Mobility seemed to be a source of considerable disagreement between the raters and the Standard. Here, ratios extended from a low of 21% in the case of patient "two" to 100% or total agreement for patient "three".

The Total Status score of the Barthel Index is a composite of the variables self care, continence, and mobility. Agreement ratios for this variable are, therefore, strongly affected by the three subsections of the scale which can often produce a masking effect. However, the Total Score of the Barthel Scale is important in the evaluation of overall treatment effectiveness. Furthermore, the Total Score of the

Barthel Index has frequently been used in the classification of patients for placement. For these reasons, it was deemed important in this study to determine the overall variation between the raters and the Standard using the complete score.

Concordance between the raters and the Gold Standard in determining functional status was lowest for patient "five" at 17% while a 96% agreement was reached for patient "three". In the second video testing of the Barthel Scale, the agreement ratios remained basically unchanged. It was the mobility status which continued to be the main factor responsible for low rater agreement with values ranging from 21% to 96%.

Table 4-1

Agreement Ratios Between 23 Raters and the Gold Standard for Two Video Sessions Using the Barthel Index (100%)

Barthel Scale						
Video I						
Patients	1	2	3	4	5	6
Self-Care	1.00	0.92	1.00	0.75	0.50	0.83
Continence	1.00	1.00	0.96	1.00	0.83	1.00
Mobility	0.54	0.21	1.00	0.63	0.29	0.38
Total	0.54	0.21	0.96	0.54	0.17	0.38
Video II						
Self-Care	1.00	0.92	1.00	0.88	0.67	0.71
Continence	0.96	1.00	0.96	1.00	0.92	1.00
Mobility	0.58	0.21	0.96	0.58	0.38	0.29
Total	0.54	0.21	0.92	0.50	0.29	0.21

Table 4-2

Agreement Ratios Between 23 Raters and the Gold Standard for Two Video Sessions Using the LORS (100%)

LORS

Video I

Patients	1	2	3	4	5	6
Home	0.42	0.33	0.38	0.25	0.33	0.79
Outside	0.25	0.08	0.46	0.38	0.29	0.38
Social	0.51	0.83	0.96	0.21	0.17	0.21
Total	0.17	0.17	0.17	0.04	0.04	0.04

Video II

Home	0.54	0.33	0.29	0.25	0.38	0.79
Outside	0.38	0.25	0.46	0.38	0.38	0.33
Social	0.54	0.83	0.88	0.08	0.17	0.38
Total	0.42	0.21	0.13	0.04	0.08	0.08

The percentage of agreement was then calculated using the Level of Rehabilitation Scale (Table 4-2). Ratios obtained for the home activities variable indicated a marked discordance between the raters and the Standard with values ranging from 25% to 79% agreement. When patients were examined in relation to outside activities, once again, it was the functional status of patient "two" that presented the greatest disagreement among the participants. Ratios were consistently low with values extending from 8% to 46%.

The variable of social activities was another source of fluctuating agreement among the twenty-three data collectors. As previously seen, it was patient "five" who posed the greatest uncertainity (17%), while in constrast, good to excellent concordance (83% and 96%) was demonstrated for patients "two and three". Like the Barthel Index, the Total Status variable of the LORS was a composite score of home, outside activities and social interaction. In general, very poor agreements between the raters and the Standard were established for all six patients with values ranging from 4% to 17%.

Repeated testing of the same patients, two weeks later, resulted in very similar ratios with the exception of patient "one" where rater agreement had improved to 42% for the Total Status variable. In addition, there was a threefold increase in agreement for patient "two" in outside activities, while agreement in social interaction for patient "four" was reduced to 8%.

In sum, the percentage of agreement ratios was useful in

providing an estimate of the concordance among the 23 raters and the independent Standard. However, this information was limited as it could only establish a trend or a relatedness of raters scores. One of the major disadvantages of this procedure was that it considered perfect ageement only and ignored the extent of agreement expected by chance alone. Furthermore, the degrees of partial agreement and disagreement were not employed in the analysis.

Measurement Bias

In appendices five through seven, Barthel Index scores recorded by the twenty-three raters are displayed along with the actual values reported by the Gold Standard for the six patients over two video sessions. On average, Group I (Evaluators) underestimated the functional status of the six patients. Nevertheless, this negative bias was mild with values ranging from -.01% to -8% (Appendix 5). One exception was seen in the case of patient "five" where a -13.3% underestimation was present in the first testing session. However, in the second testing of the Evaluators, this negative bias was reduced to -9%.

The Interpreters (Group II) were also inclined to underestimate the six patients' functional status (Appendix 6). As seen previously with the Evaluators, the negative bias was relatively small with values ranging from -1.8% to -7%. Once again, it was the fifth patient which presented the greatest variation between the Interpreters and the Gold Standard (-11.4% Video I and -7.4% Video II). However, for patient "one and two", in particular, there was a total absence of group bias between the Interpreters and the Standard.

In constrast to the first two groups of raters, the four Instructors demonstrated a slight overestimation of the functional status of the six patients for both testing sessions (Appendix 7). The values ranged from +.002% to +4.4%. The major exception was again patient "five" with a

negative bias of -27% when compared to the Standard. However, this negative bias was completely eliminated in the repeated testing of the Instructors, two weeks later. On average, the individual Instructors agreed more frequently with the Gold Standard thus reducing the group's collective bias. For all groups, individual rater bias was clearly apparent, nevertheless, no single rater consistently measured below or above the group mean.

The three separate tables examining group bias were then collapsed to estimate the overall group bias and comparisons were made with the Gold Standard (Table 4-3). On average, the raters' underestimation of functional status in the first testing session was non consequential for all practical purposes. The one exception was again patient "five" where a 17% negative bias was present. However, in the set of scores from the second video session, there was a marked decrease in group bias (-.003% to -5.3%). In sum, it was the Evaluators (Group I) who showed the greatest discrepancy using the Barthel Index when compared to the Standard. In general, they underestimated the functional status of the six selected patients while the Instructors (Group III) were relatively compatable with the norm. Interestingly, the Interpreters' scores continued to fall between the other two groups of raters. Nevertheless, when difficult or ambiguous subjects were introduced, all groups tended to lean towards the underrating of patients' performance. In the second testing of rater reliability, group bias was consistently reduced for all groups.

Table 4-3

Differences between the Gold Standard and all Raters using the Barthel Index Scores in Two Video Sessions

	Patients					
	1	2	3	4	5	6
Gold Standard	95	84	80	99	89	94
Group I (Evaluators)						
VideoI	92.9	84.7	80.0	91.0	77.1	89.7
VideoII	93.2	82.3	79.2	94.7	81.1	89.1
Group II (Interpreters)						
VideoI	96.8	87.5	78.0	97.2	78.8	94.2
VideoII	95.0	89.8	80.0	97.0	82.4	88.8
Group III (Instructors)						
VideoI	95.0	87.7	80.0	99.2	65.4	95.2
VideoII	98.7	87.7	80.0	99.0	89.0	97.7
Overall Group Means						
VideoI	94.9	86.6	79.3	95.8	73.7	93.0
VideoII	95.6	86.6	79.7	96.9	84.1	91.8
Overall Group Bias *						
VideoI	-.10	2.6	-.66	-3.2	-15.2	-0.96
VideoII	.63	2.6	-.26	-2.1	-4.80	-2.10
Percentage of Difference						
VideoI	-.001	3.1	-.008	-3.2	-17.0	-1.0
VideoII	.006	3.0	-.003	-2.1	-5.3	-2.2

* Overall Group Bias is the measure of the difference between all raters scores and the true scores (Gold Standard)

Similar information is presented for the Level of Rehabilitation Scale(Table 4-4). Once again, there was a tendency among all raters to underestimate the functional status of the six patients. In the first testing session, the Evaluators (Group I) demonstrated a negative group bias in four of the six patients assessed with a percentage of difference ranging from -4.6% to a high of -22% when compared to the Gold Standard (Appendix 8). For the remaining two patients (subjects one and six) there was a noted overestimation of the status with values of +6.7% to +18.2%.

Retesting the LORS two weeks later, the Evaluators increased their range of negative bias from a -6.5% to -26% difference with the the Gold Standard for the same four subjects. However, the reported overestimation of patients was markedly reduced to only +1% for subject two and +5.2% for the sixth subject.

The Interpreters (Group II) followed a similar pattern as the Evaluators in using the LORS (Appendix 9). A negative bias was present in the same four patients. The underestimation of status was the least for patient "three" at -2% and the greatest difference was recorded again for patient "four" at -23%. For the remaining two subjects, the Interpreters followed the same trend as the Evaluators in underestimating the patients' function, however, the bias was noticeably increased. There was a 9% positive bias in the case of patient "six" while there was a +37% overestimation for patient "two". In the second testing session of the Interpreters, the negative patients was clearly reduced.

The Instructors as a group demonstrated the highest negative bias for patient four (31%) when compared to all other raters and the Standard (Appendix 10). In addition, there was marked underestimation of function for patient "five", -21.8%. The range of positive bias was however in keeping with the other two groups. For the retesting of the LORS, the Instructors continued to show -13% to -24% underestimation of status for patients "four and five" while the positive bias was increased to 15.6% for patient "six". One major change in the Instructors decision-making was in the classification of status for patient "two". Here, the Instructors fluctuated between a noted overestimation in the first testing session, +18.7% to a slight underestimation of -2.6% for the same subject in the second video session.

When the differences between all raters and the Gold Standard were examined for the LORS, the trend towards underestimation was still prominent. Clearly, patients "four and five" drew the greatest negative bias (-25% and -17% respectively). The second patient on the other hand, was classified by all raters as having a higher level of function than that determined by the Gold Standard. In examining the second set of scores, the magnitude of the negative bias persisted (patient four, -25% and patient five, -14.7%) however, the level of positive bias was strikingly reduced.

From calculating the measurement bias for both the Barthel Index and the LORS, preliminary estimates of individual and group variations with the Gold Standard could be established. In addition, the direction and the percentage

of rater deviation Standard were determined, however, the sources of rater variation and the magnitude were yet to be measured.

Table 4-4

Differences between all raters and the Gold Standard in using the LORS in Two Video Sessions

	Patients					
	1	2	3	4	5	6
Gold Standard	78	28	67	57	63	40
Group I (Evaluators)						
VideoI	74.3	33.1	63	44.3	54.0	42.7
VideoII	72.9	28.0	59	42.1	52.8	42.1
Group II (Interpreters)						
VideoI	72.4	38.2	65.6	43.6	53.0	43.4
VideoII	70.4	29.8	66.0	42.8	53.6	41.4
Group III (Instructors)						
VideoI	77.7	33.2	68.7	39.2	49.2	42.5
VideoII	78.0	27.2	67.5	43.2	54.5	46.2
Overall Group Means						
VideoI	74.8	34.8	65.7	42.3	52.0	42.8
VideoII:	73.7	28.3	64.1	42.7	53.6	43.2
Overall Group Bias *						
VideoI:	-3.2	6.80	-1.2	-14.6	-10.7	2.8
VideoII:	-4.2	.33	-2.8	-14.3	- 9.9	3.2
Percentage of Difference						
VideoI	-4.1	24.0	-1.7	-25.6	-17.0	7.0
VideoII	-5.3	1.1	-4.1	-25.0	-14.7	8.0

* Group Bias is the measure of difference between all rater scores and the true scores (Gold Standard)

The Evaluation of the Variation between the Raters and the Gold Standard through the Analysis of Variance

To identify the sources and magnitude of the variability among the three groups of raters and the Gold Standard, a repeated measures analysis of variance (ANOVA) for a mixed model was employed. The total variance for each of the functional subscales was estimated, then divided into five separate components and compared to the Gold Standard. The sources of variation were attributed to: 1) the differences among the three groups of raters, 2) the differences among the raters within the same group, 3) the differences among the patients themselves, 4) the differences between individual raters and specific patients within the same group, and 5) random error. For this study, the error component consisted of error due to the video sessions themselves plus random error.

From the Analysis of Variance, the Expected Mean Squares (EMS) and the Coefficients of Variation (CV %) could be obtained. By definition, the Expected Mean Squares are the estimate of variance attributed to each of the components of an equation. The coefficient of variation is the ratio of the Standard deviation to the mean.

The Barthel Index was examined in depth through the subscales of Self Care, Continence, Mobility as well as by looking at the Total Status to determine the percentage of variability among the raters in in comparison to the Gold Standard. The greatest source of variation for the subscale

of Self Care was credited to the difference between specific raters and particular patients within Group I and Group II (Appendix 11). A 4.2% variation existed between the Evaluators and the Standard while a 5% variation was present for the Interpreters. The Instructors (Group III), on the other hand, were in complete agreement with the Standard (0% variation). All other sources of variation for the Self Care subscale were minimal.

A similar pattern of variation break down was demonstrated for the subscale of Continence (Appendix 12). Once again, the greatest source of variation was assigned to rater-patient differences for Group I (4.4%) and Group II (7.5%). In general, the Interpreters showed more variability than the other two groups when using the Continence subscale, yet the coefficients of variation were less than 4% for each remaining causes of variation.

In constrast to the first two subscales, patient variation was the main source of variability for the subscale of Mobility. (Appendix 13). As a group, the Evaluators were more inclined to differ with the Gold Standard (9.5%). There also was considerable within raters variation for all three groups (Evaluators 7.6%; Interpreters 4%, and Instructors 2.8%). Furthermore, the random error component was elevated for this subscale.

The overall variation among the raters and the Standard was also tested using the composite value of the three subscales which gave the Total Status score for the Barthel Index (Table 4-5). Other than patient differences, the

coefficients of variation were most pronounced for individual rater-patient differences for all groups. However, it was Group I, (Evaluators) who continued to demonstrate the greatest variability in using the Barthel Index when compared against the Standard.

The expected mean squares and the coefficients of variation were then calculated for all raters using the Level of Rehabilitation Scale. Like the Barthel Index, the LORS was divided into subscales of home activities, outside activities, social interaction and were summed to produce the total status score. However unlike the Barthel Index, the Level of Rehabilitation Scale appeared to be a source of increased variation for the raters when compared to the Standard.

For the subscale of Home Activities (Appendix 14), the individual patient differences accounted for better than 30% of the recorded variation within each group. Rater-patient interaction was the next major variance component, with the Evaluators and the Instructors demonstrating higher coefficients of variation (12.6% and 12.9%) than that recorded by the Interpreters (9.4%). Individual differences within each of the groups of raters was less than 10%, nevertheless, it was the Instructors who demonstrated greater discrepancy with the Standard. In addition, there was discordance among the groups of raters but this was present only for the Evaluators and the Interpreters (6.1% and 5.9%).

The second subscale of the LORS to be tested was that of Outside Activities (Appendix 15). Once again, better than 45%

of the variation for all groups of raters was attributed to differences among the patients themselves. Deviations among certain raters and patients continued to prevail, nevertheless, it was the Evaluators who demonstrated the greatest inconsistencies with the Standard (23%). An 11.4% discrepancy was also present for the Evaluators' "within group" component, whereas zero variability was recorded for both the Interpreters and the Instructors. In constrast to all other subscales, there was a total absence of group variation with the Gold Standard.

In general, there appeared to be a greater individual as well as group understanding in the use of the Social Interaction subscale (Appendix 16). The main source of variation was again the patients themselves with specific rater-patient variation being the other major cause of discrepancy. Unlike the previous subscales, there was no individual rater variation present. Group variation was again non-existent in the recording of social activities.

As in the case of the Barthel Index, the Total Status component of the LORS was examined for the overall variation of the three groups in comparison to the Gold Standard (Table 4-6). Generally, there were strong similarities in the spread of variation for each of the groups of raters. However, individual raters within the Evaluator group continued to demonstrate slightly greater deviations when the Level of Rehabilitation Scale was used. In order to interpret the magnitude of this variation, coefficients of reliability were then calculated.

Table 4-5

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Total Status Section of the Barthel Index

Barthel Index

Total Status

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Group			Group		
	I	II	III	I	II	III
+ Var(groups)	8.1	0.00	0.0	3.2	0.0	0.0
++ Var(rater within group)	7.5	5.4	1.2	3.2	0.0	0.0
* Var(patients)	39.1	53.20	54.3	7.2	8.2	8.1
** Var(rater x patient within group)	15.6	8.80	3.0	4.5	3.3	1.9
*** Var(video+ random error)	0.6	0.01	3.1	0.85	0.1	1.9

* var(groups)=variation between groups
++ var(r in group)=variation of rater in group)
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

Table 4-6

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Total Status Section of the LORS

LORS

Total Status

Sources of Variation	Expected Mean Squares Group I	Group II	Group III	Coefficient of Variation Group I	Group II	Group III
+ Var(groups)	9.9	16.6	5.2	6.1	7.8	4.3
++ Var(rater within group)	14.6	0.0	0.0	7.4	0.0	0.0
* Var(patients)	246.1	258.3	327.4	30.6	31.0	34.0
** Var(rater x patient within group)	23.3	23.3	24.1	9.4	9.3	9.3
*** Var(video+ random error)	4.1	0.34	0.53	3.9	1.1	1.3

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group)
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

Level of Concordance among the Raters and the Gold Standard

Taking the expected mean squares from each of the subscales the Intra-Class Correlation Coefficients (ICC) were estimated between the three groups of raters and the Gold Standard (Table 4-7). Coefficients for the Barthel Index ranged between R=0.00, a complete absence of agreement to the maxium value of R=+1.00, total concordance. It was the Continence subscale that produced the greatest disagreement for the Evaluators and the Interpreters when compared to the Standard (ICC=0.00 Evaluators and ICC=0.12 Interpreters). Similarly, there apparently was no agreement between the Interpreters and the Standard for the activities of self care(ICC=0.02). The agreement ratio between the Evaluators and the Standard was also extremely low by most criteria, however, test of consistencies for this subscale proved to be statistically significant. In constrast, the Instructors were shown to be in complete agreement with the Gold Standard for these two subscales. On average, all groups of raters demonstrated strong coefficients of agreement for the Barthel subscales of Mobility and Total Status. As estimated from the examination of group variations, the Evaluators had the lowest agreement levels for these subscales when compared with the Standard (R=0.65and R=0.55 respectively). The Interpreters and the Instructors, on the other hand, had similar intra-class correlations for Mobility and Total Status and were in stronger concordance with the Standard.

The Level of Rehabilitation Scale resulted in good to excellent agreements among the three groups of raters and the Standard with correlations ranging from R=0.76 to R=0.96 (Table 4-7). The lowest agreement ratio (R=0.76) came from the Evaluators' interpretation of Outside Activities. As reported earlier, this particular subscale was the source of increased variation for the Evaluators which can explain the lower agreement ratio for this group. In all cases, however, the intra-class correlation coefficients were stochastically significant. Although the quantitative significance of the ICC usually depends on its own magnitude (Kramer and Feinstein, 1981), routine tests of consistence were performed for all subscales in this study in an attempt to compare the degrees of rater variation recorded with the groups' level of concordance.

Table 4-7

Overall Agreement Between each Group of Raters and the Gold Standard as Measured by the Intra-Class Correlation Coefficient (ICC) for the Barthel Index and the LORS

Barthel Index

Dependent Variables	Gold Standard: Evaluators+ N=180 (Sig)	Interpreters+ N=72 (Sig)	Instructors+ N=60 (Sig)
Self Care	0.20 (0.0001)	0.02 (0.36)	1.00+
Continence	0.00 (0.47)	0.12 (0.11)	1.00+
Mobility	0.65 (0.0001)	0.87 (0.0001)	0.88 (0.0001)
Total Status	0.55 (0.0001)	0.79 (0.0001)	0.88 (0.0001)
LORS			
Home Activities	0.82*	0.87*	0.80*
Outside Activities	0.76*	0.89*	0.93*
Social Activities	0.92*	0.95*	0.96*
Total Status	0.83*	0.87*	0.92*

+ Complete Agreement
* $p<0.0001$)

Table 4-8

Tests of Conformity for the Barthel Index and the LORS Scores When Comparing the Three Groups of Raters and the Gold Standard

Barthel Index

Dependent Variables	Gold Standard + Evaluators (df=1,70)	+ Interpreters (df=1,25)	+ Instructors (df=1,20)
	F value (Sig)	F value (Sig)	F value (Sig)
Self Care	2.61 (0.11)	1.40 (0.24)	0.00 (---)
Continence	0.31 (0.58)	0.78 (0.38)	0.00 (---)
Mobility	3.65 (0.06)	0.01 (0.97)	0.62 (0.44)
Total Status	5.20 (0.02)	1.03 (0.31)	0.62 (0.44)
	LORS		
Home Activities	3.54 (0.06)	4.62 (0.04)	0.86 (0.36)
Outside Activities	0.25 (0.61)	0.11 (0.73)	0.51 (0.48)
Social Activities	0.19 (0.66)	0.39 (0.53)	1.06 (0.31)
Total Status	5.05 (0.02)	4.48 (0.04)	1.80 (0.19)

To assess the presence of any systematic bias between the three groups of raters and the Gold Standard, tests of conformity were tabulated (Table 4-8). Statistically significant bias was found between the Evaluators and the Standard for the Total Status measure for both the Barthel Index and the LORS (F value=5.20; $p<0.02$ and F value=5.05, $p<0.02$). Similarly, systematic bias was present between the Interpreters and the Standard for the LORS subscales of Home Activities and Total Status (F value=4.62; $p<0.04$ and F value=4.48; $p<0.04$). The Instructors, on the other hand, were either in perfect agreement or demonstrated high levels of conformity with the Standard for both Scales. As previously reported, there appeared to be a trend towards the underestimation of patient performance by certain raters. This descriptive underestimation of rater scores was simply formalized when tests of conformity were performed.

Comparison Among the Three Groups of Raters

The next procedure was to examine the inter-rater reliability among the raters themselves. In Table 4-9, the means, Standard deviations and the range of scores were compared for each of the six patients interviewed in two distinct video testing sessions. When the Barthel Index was used there appeared to be an overall consensus among the 23 raters in determining patient levels of functional status. Rater variability was present, nevertheless, with the greatest deviations being recorded by the Evaluators (Group I) and the least discrepancy being posted by the Instructors (Group III). For the most part, the Interpreters' scores (Group II) fell midway between the other two groups of raters. An additional point of interest was that the all six patients were rated above the 75th percentile on the Barthel Index.

A similar pattern of scores was seen in the second testing of the Barthel Index. Once again, it was the Evaluators who varied the most in establishing the levels of function for each patient. However, all groups had difficulty in scoring patient "five" (Group I= 81.3 ±9.1; Group II=82.0 ±8.2; Group III=89.0 ±7.1) while near perfect agreement was reached for patient three (Group I=79.2 ±2.6; Group II=80 ±0; Group III=80 ±0) (Table 4-9).

The Level of Rehabilitation Scale (LORS) was assessed in a similar manner (Table 4-10). Although there was a wide range of functional levels among the six patients, the mean

values recorded by the three sets of raters were within five to six points of one another. In comparison to the Barthel Index, however, there was an overall increase of "within group" variation as well as a gain in the range of scores for each cohort of raters.

When the second testing of the LORS was examined the raters' scores appeared to be compatible for the two sessions. There was better precision in the "within group" agreement but inconsistencies were still present for particular raters.

Table 4-9

Mean Scores SD and (Range) recorded by the Three Groups of Raters for Six Selected Patients in Two Video-Testing Sessions using the Barthel Index

Mean Barthel Scores

Video I

	Patients					
	1	2	3	4	5	6
GroupI	92.8 5.1 (80-100)	84.1 6.0 (74-90)	80 0 (80-80)	91.1 9.1 (71-99)	77.1 8.3 (63-91)	89.4 7.2 (73-99)
GroupII	96.8 2.9 (94-100)	87.2 3.0 (83-90)	78.0 4.4 (70-80)	96.8 3.0 (93-99)	78.8 8.6 (68-89)	94.4 .9 (94-96)
Group III	95 0 (95-95)	87.7 2.5 (84-89)	80 0 (80-80)	99.3 .1 (99-100)	81.8 4.8 (79-89)	95.3 2.5 (94-99)

Video II

GroupI	93.2 3.7 (85-100)	82.4 6.8 (74-90)	79.2 2.6 (70-80)	94.7 4.3 (84-100)	81.3 9.1 (67-94)	89.1 6.9 (79-99)
GroupII	95.0 3.5 (90-100)	89.8 3.5 (85-95)	80 0 (80-80)	97.0 2.7 (94-100)	82.0 8.2 (69-89)	88.8 7.4 (78-99)
GroupIII	98.7 2.5 (95-100)	87.7 2.5 (84-89)	80 0 (80-80)	99 0 (99-99)	89.0 7.1 (79-94)	97.7 2.5 (94-99)

Table 4-10

Mean Scores, SD, and (Range) recorded by Three Groups of Raters for Six Selected Patients in Two Video-Testing Sessions using the Level of Rehabilitation Scale

Mean LORS Scores

Video I

	Patients					
	1	2	3	4	5	6
GroupI	74.4 6.9 (59-88)	33.2 11.1 (25-68)	62.9 9.2 (44-77)	44.3 7.8 (32-58)	54.1 6.5 (41-62)	42.7 6.3 (31-53)
GroupII	72.4 8.1 (62-82)	30.6 4.7 (23-35)	65.6 4.5 (59-70)	43.4 4.4 (38-50)	53.0 6.5 (41-62)	43.4 6.3 (38-47)
GroupIII	78.5 5.1 (71-82)	33.2 10.3 (23-46)	69.0 3.4 (64-72)	39.5 7.1 (32-47)	49.2 5.1 (44-59)	42.5 3.8 (38-47)

Video II

GroupI	74.4 6.4 (56-79)	28.1 3.3 (23-34)	59.1 10.0 (41-73)	42.1 6.3 (32-53)	52.8 7.4 (41-63)	42.1 4.7 (31-50)
GroupII	70.6 8.3 (59-79)	29.8 8.4 (20-43)	66.0 4.9 (60-72)	42.8 1.6 (41-44)	53.6 7.1 (47-65)	41.4 4.8 (34-47)
GroupIII	78.5 3.3 (74-82)	27.2 5.7 (20-34)	67.7 5.5 (60-73)	43.2 4.5 (38-47)	54.5 1.7 (53-56)	46.2 5.1 (41-53)

Correlations Between the Three Groups of Raters

In this section of the analysis, each rater cohort was examined collectively and comparisons of the instrument scores were estimated using first the Product Moment Correlation. Associations among the three groups of raters for the Barthel Index are presented in Table 4-11. A total of 228 patient functional status scores were examined between the Evaluators and the Interpreters. Although statistically significant correlations were present for the subsections of Self Care ($r=0.20$, $p<0.05$), Mobility ($r=0.33$, $p<0.0001$), and Total Status ($r=0.26$, $p<0.01$), the coefficients were relatively low. Moreover, there was no association seen between the groups for the status for the Continence subscale ($r=.02$).

The Evaluators and the Instructors were then compared on a total of 216 Barthel Scores. As seen in the first group comparison, there was again no relationship between the groups for the Continence subscale ($r=0.01$), while moderate statistical significant correlations were once again established for the status of Self Care ($r=0.24$, $p<0.05$), Mobility ($r=0.26$, $p<0.01$), and Total Status ($r=0.24$, $p<0.05$).

The third collate compared the Interpreters and the Instructors in a sample of 108 scores. The subscales of Self Care and Continence continued to be the major sources of inconsistency among the raters ($r=0.02$ and $r=0.03$), while, the subscales of Mobility ($r=0.24$, $p<0.01$) and Total Status ($r=0.22$, $p<0.05$) followed a similar pattern seen in the previous group comparisons.

The next step was to examine the correlations among the three groups using the Level of Rehabilitation Scale (LORS) (Table 4-11). In constrast to the Barthel Index, moderate to good associations were present between the paired groups of raters for each of the activities examined. Although the correlation coefficients tended to be low, statistical significance was obtained in all cases. The measure of association between Group I and Group II ranged from r=0.45 for Home Activities to r=0.71 for the Total Status subscale. Statistical significance for each subscale was recorded at $p<0.0001$.

Stronger measures of associations were seen between the Evaluators and the Instructors for the same four subscales. The lowest relationship was obtained for social interaction ($r=0.63$, $p<0.0001$), while the remaining three components of the LORS had stronger relationships with values ranging from $r=0.73$, ($p<0.0001$) to $r=0.84$, ($p<0.0001$).

Like the first group comparison, the Interpreters and the Instructors had relatively low levels of associations for the same four subscales of the LORS ,nevertheless, all measures proved to be statistically significant ($r=0.66$, $p<0.0001$ to $r=0.72$, $p<0.0001$).

The Pearson correlation coefficient has traditionally been the measure of choice for assessing observer concordance. However, the product-moment correlation coefficient is a bivariate statistic which can only determine the relationship between two variables (Hasselkus, 1976). These correlation

indices have the advantage of considering both the chance-expected agreement and partial agreement and disagreement. However, they are a measure of linear relatedness and not of concordance or sameness. As a result, this coefficient can only give an indication of the association or trend of scores and completely ignores any systematic bias that may exist between different raters. Thus, if the task to be examined is a repeated measure or if more than one rater is involved, only one variable is being measured and the bivariate statistic should not be employed. Therefore, the Analysis of Variance has now become the method of choice. From this maneuver, the estimates of variance can be determined which are then used to compute the Intra-class Correlation Coefficients.

Table 4-11

Product Moment Correlations Between the Groups of Raters Using the Barthel and LORS in Two Testing Sessions for Six Patients

	Evaluators and Interpreters (N=228)	Evaluators and Instructors (N=216)	Interpreters and Instructors (N=108)
Barthel Index			
Self Care	0.20*	0.24+	0.02
Continence	0.02	0.01	0.03
Mobility	0.33**	0.26+	0.24+
Total Barthel	0.26+	0.22*	0.22*
LORS			
Home Activities	0.45++	0.73++	0.72++
Outside Activities	0.60++	0.73++	0.72++
Social Activities	0.62++	0.63++	0.68++
Total LORS	0.71++	0.84++	0.66++

* $p<0.05$
\+ $p<0.01$
** $p<0.001$
++ $p<0.0001$

Group I =evaluators
Group II =interpreters
Group III=instructors

Based on this reasoning, the inter-observer reliability was assessed using the Analysis of Variance (Anova). The Expected Means Squares and the Coefficients of Variation were again estimated for the subscales of both the Barthel Index and the Level of Rehabilitation Scale (LORS). As seen in the rater-Standard comparisons, the observer-patient component was the greatest source of variation for the Barthel subscale of Self Care (Appendix 17). The variation between certain raters in specific groups was greater than the variation due to differences in the groups themselves. On the other hand, the variance attributed to the video-testing sessions and random error was negligible. The lack of patient variability was again evident for this subscale.

The subscale of Continence followed a similar pattern as the preceeding scale with the "rater-patient within a group" variance being the most prominent contributor (Appendix 18). The values ranged from 4% for the Evaluator-Instructor comparison to 6.2% for the Interpreter-Instructor combination. Overall, the presence of variation was slight for this subscale.

For the subscale of Mobility, there appeared to be a better agreement between the groups and the Gold Standard than when the groups were compared alone. Nevertheless, the percentage of variation was less than 10% in both comparisons. The primary source variability for this subscale was between the patients themselves which permitted a clearer assessment of the actual inter-rater reliability for these groups

19).

In general, the individual group variances for the Total Status subscale were comparable to the group-Standard comparison. The 5% variation between the Interpreters and the Instructors was the most prominent inconsistency among the groups of raters (Table 4-12).

The three groups of raters were next compared using the Level of Rehabilitation Scale (Appendix 20-22; Table 4-13). Suprisingly, there was very little variation present among the data collectors for all subsections of this scale. The one exception was in the case of the Home Activities component. In this instance, the Group II-Group III comparison registered a 2.8% variation. In order of magnitude, the principle sources of variation for all sections of the scale, excluding patient differences, were the "rater-patient" and "rater within a group" components. There appeared to be a slightly greater video and random error variation when using the LORS which was similar to the group-Standard comparison.

Table 4-12

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Total Status Score of the Barthel Index

Barthel Index

Total Status

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
		Groups			Groups	
	I + II	I + III	II + III	I + II	I + III	II + III
+ var(groups)	3.50	0.0	19.16	2.1	0.0	5.0
++ var(rater within group)	6.70	2.68	5.78	3.0	1.8	2.7
* var(patients)	43.40	56.25	42.49	7.5	8.3	7.4
** var(rater x patient within group)	14.59	6.99	13.99	4.3	2.9	4.2
*** var(video+ random error)	0.37	1.07	1.23	0.7	1.1	1.2

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group)
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

Table 4-13

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Total Status Score of the LORS

LORS

Total Status

Sources of Variation	Expected Mean Squares Groups I + II	I + III	II + III	Coefficient of Variation Groups I + II	I + III	II + III
+ var(groups)	0.00	0.00	0.00	0.0	0.0	0.0
++ var(rater within group)	15.29	15.33	3.94	7.6	7.6	3.8
* var(patients)	238.36	254.84	273.40	30.3	31.0	32.0
** var(rater x patient within group)	21.70	23.18	20.12	9.1	9.3	8.6
*** var(video+ random error)	3.13	2.16	0.0003	3.4	2.8	0.3

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group)
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

Coefficients of Reliability among the Three Groups of Raters

The estimates of variance were then used to compute the intraclass correlation coefficients (ICC) between the groups (Table 4-14). The Barthel Index continued to generate the lowest levels of agreement among the raters for the first two subscales and good to excellent ratios for the remaining sections. The coefficients for the activities of Self Care varied from R=0.01 (no agreement) between the Interpreters and the Instructors to R=0.25, ($p<0.0001$) for the Evaluators and the Instructors. The main reason for the poor concordance among the groups was again centered around the Continence subscale. However, better agreement was evident for the sections of Mobility and Total Status. Once again, it was the Interpreters and the Instructors who demonstrated the strongest agreement ratios (Mobility, R=0.90; Total Status, R=0.80). Levels of concordance continued to range from good to excellent among all raters when the LORS was used. The lowest value was recorded by the Evaluators and the Interpreters for the subscale of Outside Activities (R=0.78) while the Interpreters and the Instructors were credited with the highest level of agreement for the section of Social Activities (R=0.97).

Although tests of conformity are normally used to compare concordance values against a selected Standard, these tests were also conducted for the raters to determine if consistency of the recorded scores also meant conformity. As seen in Table 4-15, a marked bias was noted between the Evaluators and

the Interpreters when the subscales of Mobility ($F=15.6$, $p<0.0002$) and Total Status ($F=8.97$, $p<0.003$) were tested. Likewise, bias was present in the Evaluator-Instructor comparison for the activities of Self Care ($F=9.75$, $p<0.0002$); Mobility ($F=21.9$, $p,0.0001$) and Total Status ($F=27.86$, $p<0.0001$). Lack of conformity was also evident between the Interpreters and Instructors for the subscales of Self Care ($F=5.56$, $p<0.02$) and Total Status ($F=8.74$, $p<0.005$).

In constrast, there was a total absence of systematic bias among the raters when the Level of Rehabilitation Scale was tested. From these results, the Barthel Index was again the scale which seemed to produce the greatest indecision among the raters even though statistical significant consistency was present in several of the subscales.

Table 4-14

Inter-Observer Reliability Between each of the Three Groups of Raters as Measured by the Intra-Class Correlation Coefficient (ICC) for the Barthel Index and the LORS

Barthel Index

Subscales and Total Scale	Evaluators and Interpreters N=228	Evaluators and Instructors N=216	Interpreters and Instructors N=108
Self Care	0.25*	0.17*	0.01 (0.36)
Continence	0.07 (0.03)	0.00 (0.46)	0.07 (0.11)
Mobility	0.66*	0.62*	0.90*
Total Status	0.67*	0.51*	0.80*
	LORS		
Home Activities	0.85*	0.84*	0.86*
Outside Activities	0.78*	0.78*	0.89*
Social Activities	0.92*	0.93*	0.97*
Total Status	0.85*	0.86*	0.92*

*p<0.0001

Table 4-15

Tests of Conformity for the Barthel Index and LORS Scores in Comparing the Three Groups of Raters

Barthel Index

Subscales and Total Scale	Evaluators and Interpreters (df=1,90)	Evaluators and Instructors (df=1,85)	Interpreters and Instructors (df=1,40)
	F value (Sig)	F value (Sig)	F value (Sig)
Self Care	0.06 (0.80)	9.75 (0.0002)	5.56 (0.02)
Continence	2.36 (0.12)	1.22 (0.27)	2.95 (0.09)
Mobility	15.63 (0.0002)	21.90 (0.0001)	1.86 (0.18)
Total Status	8.97 (0.003)	27.86 (0.0001)	8.74 (0.005)
	LORS		
Home Activities	0.09 (0.75)	1.99 (0.16)	1.04 (0.31)
Outside Activities	0.18 (0.67)	0.00 (0.96)	0.13 (0.39)
Social Activities	0.06 (0.80)	0.91 (0.34)	0.74 (0.39)
Total Status	0.03 (0.87)	1.90 (0.17)	1.31 (0.25)

Comparison of Within Rater Reliability

The next step was to examine the repeated measurement scores for each of the groups of raters to estimate the intra-rater reliability. Two weeks after the initial evaluations, all raters were required to re-assess the six patient video-taped interviews. It was felt that the taped sessions would eliminate any temporal change in patients' status and reduce problems of patient learning and fatigue. The within group means and standard deviations for the Barthel Index are presented in Table 4-9.

On average, it was the Evaluators who again showed the greatest variability over time, although the overall variation was slight. There was some degree of fluctuation from time one to time two for all of the recorded scores, the most noted being seen for patient "three and four". In general, there was a decrease in variation among the Interpreters within this time frame with the exception of patient "six". Here, the four Interpreters as a group were clearly uncertain as to the level of ability for this patient in the second evaluation session. The Instructors, on the other hand, were relatively consistent over time with only a slight increase in recorded scores being seen for patient "one and five".

The means and the standard deviations for the LORS were also examined in a similar manner (Table 4-10). In constrast to the Barthel Index, the most noted source of variability stemmed from the group of Instructors while the Interpreters showed the least variation over this two week period. The

Instructors reported higher levels of function for the first set of patient evaluations in comparison to the other two groups of raters but reduced these levels of function during the second scoring session. As seen previously, patient "two" caused the greatest indecision among all raters when the LORS was the instrument being tested.

It is rare, however, that a series of scores are exactly the same. In general, they tend to vary. It is this source of variability that is important to the reliability of the scale.

Table 4-16

Product Moment Correlations of the Test-Retest Sequence for the Three Groups Using the Barthel Index and the LORS

	Group I (N=168)	Group II (N=60)	Group III (N=48)
Barthel Index			
Self Care	0.21*	0.26	0.00
Continence	0.21*	0.29	0.00
Mobility	0.61++	0.77++	0.81++
Total Barthel	0.54++	0.73++	0.81++
LORS			
Home Activities	0.88++	0.84++	0.89++
Outside Activities	0.79++	0.91++	0.92++
Social Activities	0.82++	0.88++	0.96++
Total LORS	0.89++	0.91++	0.93++

* p<0.05
\+ p<0.01
** p<0.001
++ p<0.0001

Group I =Evaluators
Group II =Interpreters
Group III=Instructors

Product Moment Correlations were used again to determine the levels of association between the two recording sessions for each of the subscales of both the Barthel Index and the LORS (Table 4-16). For the Evaluators, the coefficients indicated that the repeated scores for the subscales of Self Care, Continence, Mobility and Total Status were indeed similar to the first test with values ranging from $r=0.21$ to $r=0.61$ ($p<0.05$ to $p<0.0001$). Statistically significant correlations were obtained for each of the four subscales but again the values were relatively low.

In the case of the Interpreters, there was an absence of significant association for the subscales of Self Care and Continence, while relatively strong associations were obtained for the components of Mobility and Total Status. Above all there appeared to be a total absence of association between the two testing sessions for the Instructor group, yet, there was a strong relationship present for Mobility (0.81, $p<0.0001$) and Total Status variables (0.81, $p<0.0001$).

The groups were then reassessed using the Level of Rehabilitation Scale (Table 4-16). For this index, the rater reproducibility appeared to be excellent. The correlations of comparison ranged from $r=0.79$ for the subscale of Outside Activities for the Evaluators to a high of $r=0.96$ for the subscale of Social Interaction as recorded by the Instructors. All correlations were significant at the 0.0001 level.

Graphs of the two video sessions were then plotted for each of the groups of raters for both the Barthel Index and

the LORS (Appendix 23 to 28). Considerable scatter was evident for each of the plots when the Barthel Index was tested. Although the subscales of this index were not represented individually, each group plot did provide a overall image of how the Barthel Scale was scored over time. The scores of each of the twenty-three raters were spread in a relatively positive direction but there was little evidence of linearity. On the other hand, the LORS plots were clearly describing the presence of a strong relationship between the two video sessions for each of the three rater groups. In all cases, the range of scores for the second video was slightly reduced.

An Analysis of Variance was calculated to examine the sources of variation within the rater groups when the six patients were re-evaluated on the two scales. The coefficients of variation are presented in Appendices 29 to 31. The Evaluator within-rater variability was examined first. As seen previously, the coefficients of variation for the four subscales of the Barthel Index were relatively low. In effect, the variance attributed to the rater-patient interaction accounted for the greatest prorportion of the total variation. The actual biological differences between the patients themselves were virtually non existant except for the subscales of Mobility (17.7%) and to a lesser extent the Total Status subscale (7.1%) (Appendix 29).

In constrast, patient differences accounted for 30% to 58% of the overall variation in the repeated use of the LORS.

There were two other areas of discrepancy: patient-rater dispariety as well as variation due to specific raters within a group. In both cases, the subscale of Outside Activities was the greatest source of rater inconsistency with values ranging 23% for rater-patient interaction and 12.8% for specific rater differences. Random error, on the other hand, continued to be minimal.

As a group, the Interpreters recorded a slightly greater variation among the six patients in using the Barthel Index over time (Appendix 30). Nevertheless, the distribution of variation followed a similar pattern to the group of Evaluators. Patient differences in the areas of Mobility and Total Status accounted for the greater proportion of the overall deviation (17.8% and 8.4%) while the largest coefficients of variation for the subscales of Self Care and Continence were less than 8%. In all cases, the rater-patient differences continued to be the next major source of overall variation while random error was equivalent to the variation attributed to the individual raters within the group.

When the LORS was the tool of assessment, the distribution of variation recorded by the Interpreters generally approximated that of the Evaluators with one exception, individual rater variation was less than 6%. Random error continued to be low with values ranging from 1.3% to 5.5%.

Unlike the other two groups, the Instructors demonstrated minimal to zero variation over time when the Barthel Index was the testing tool (Appendix 31). However, individual patient

differences were also non-existent in two of the four subscales. Rater-patient variation was present for the Mobility and Total Status subsections, yet the percentage of variation was limited to 3.7%. Furthermore, random error was greater than the "rater within the group" and "rater-patient" variations (2.4% to 5.8%).

The within-rater variation for the Level of Rehabilitation Scale followed the pattern set by the first two groups. One major difference, however, was the total absence of deviation attributed to the "rater within the group component for the subscales of Outside and Social Activities. Furthermore, 10.6% of the variation for Outside Activities was listed as random error.

Coefficients of Reliability for the Within Rater Agreement

From these estimates of variance, the intra-class correlation coefficients (ICC) were computed (Table 4-17). The overall intra-rater agreement was significant in 75% of the Barthel Index. It was the subscale of Continence which proved to be the major source of indecision for the Evaluators with a total absence of agreement noted in the repeated use of this scale. Equally, the Interpreters had difficulty with this subscale ($R=0.15$, $p<0.11$). The Instructors, on the other hand, tended to repeat the same or similar scores over the two testing sessions. The LORS continued to produce good to excellent agreement for all groups with values ranging from $R=0.74$ ($p<0.0001$) for the subscale of Outside Activities in the case of the Evaluators to a high of $R=0.98$ ($p<0.0001$) for the Instructors in Social Activities.

Table 4-17

Intra-Observer Reliability for the Three Groups of Raters as Measured by the Intra Class Correlation Coefficient (ICC) for the Barthel Index and the LORS

	Evaluators N=168 (Sig)	Interpreters N=60 (Sig)	Instructors N=48 (Sig)
	Barthel Index		
Self Care	0.30 (0.0001)	0.030 (0.37)	1.00+
Continence	0.00 (0.47)	0.15 (0.11)	1.00+
Mobility	0.67 (0.0001)	0.87 (0.0001)	0.90 (0.0001)
Total Barthel	0.50 (0.0001)	0.77 (0.0001)	0.89 (0.0001)
	LORS		
Home Activities	0.84*	0.90*	0.82*
Outside Activities	0.74*	0.86*	0.92*
Social Activities	0.91*	0.97*	0.98*
Total LORS	0.84*	0.90*	0.93*

* $p<0.0001$
+ Complete Agreement

In sum, there appeared to be excellent agreement between the raters and the Gold Standard in the scoring of the six patients when the Level of Rehabilitation Scale was the testing instrument. Good to excellent inter-rater and intra-rater reliability were also demonstrated. In constrast, the overall concordance for the first two subscales of Barthel Index ranged from poor in the case of the Interpreters to complete agreement for the Instructors and the Standard. On the other hand, acceptable levels of agreement were achieved for the Mobility subscale and Total Status.

Inter-rater reliability between the Evaluators and the Interpreters was present for the Barthel Index but with relatively low values. Similiarly, fair agreement was recorded between the Evaluators and the Instructors. The strength of the agreement was the greatest between the Instructors and the Interpreters but only for the last two sections of the scale as there was a total absence of agreement for the subscales of Self Care and Continence. In essence, the Continence subscale was poorly understood by all raters. When the within-rater reliability was examined, agreement levels resembled the pattern set between the individual groups and the Standard.

On-the-Spot Inspections

In order to verify the qualilty of the data being collected over the life of the Geriatric Study, on going inspections were conducted for the active data collectors. Simultaneous scoring of patient status were recorded by the rater assigned to the interview and one instructor for 105 Hospital Visits and 46 Home Visits. As seen in Table 4-18, strong agreements were present between the two raters for the Hospital evaluations when the Barthel Scale was used (R=0.96 for Continence and R= 0.99 for Total Status).

In the Home Setting, the Barthel and the LORS Instrument were used together to evaluate the patients' status. Again, relatively good to excellent correlations were produced which demonstrated high precision in estimating the patients' present state of function. Use of Barthel Index indicated good agreement between the raters with scores ranging from R=0.97 to R=1.00, while the LORS ranged from R=0.75 for Social Activities and R=1.00 for Total Status (Table 4-19).

Table 4-18

INTER-OBSERVER RELIABILITY BETWEEN RATERS AND INSTRUCTOR
FOR ON-THE-SPOT INSPECTIONS IN THE HOSPITAL SETTING
AS MEASURED BY THE INTRACLASS CORRELATION COEFFICIENT
(N=105)/

BARTHEL INDEX

RATER

INSTRUCTOR	SCSUBT	CONSUBT	MOBSUBT	BARTOT
SCSUBT2	0.9582 0.0001	0.7821 0.0001	0.7427 0.0001	0.9102 0.0001
CONSUBT2	0.8039 0.0001	0.9622 0.0001	0.6005 0.0001	0.8291 0.0001
MOBSUBT2	0.7119 0.0001	0.5954 0.0001	0.9618 0.0001	0.8936 0.0001
BARTOT2	0.9023 0.0001	0.8187 0.0001	0.9028 0.0001	0.9840 0.0001

Scsubt =Self Care Subscale of the Barthel Index
Consubt =Continence Subscale of the Barthel Index
Mobsubt =Mobility Subscale of the Barthel Index
Bartot =Total Score of the Barthel Index

Table 4-19

INTER-RATER RELIABILITY BETWEEN RATERS AND INSTRUCTOR
FOR ON-THE-SPOT INSPECTIONS IN THE HOME SETTING
AS MEASURED BY THE INTRACLASS CORRELATION COEFFICIENT
(N=46)

BARTHEL INDEX

		RATER			
		SCSUBT	CONSUBT	MOBSUBT	BARTOT
INSTRUCTOR	SCSUBT2	0.9707 0.0001	0.4071 0.0050	0.7219 0.0001	0.8606 0.0001
	CONSUBT2	0.3548 0.0155	0.9212 0.0001	0.4749 0.0009	0.5549 0.0001
	MOBSUBT2	0.7428 0.0001	0.3943 0.0067	0.9682 0.0001	0.9453 0.0001
	BARTOT2	0.8630 0.0001	0.5177 0.0002	0.9352 0.0001	0.9802 0.0001

Scsubt =Barthel Index Subscale of Self Care
Consubt=Barthel Index Subscale of Continence
Mobsubt=Barthel Index Subscale of Mobility
Bartot =Barthel Index Subscale of Total Score

LORS

		RATER			
		HASUBT	OASUBT	SASUBT	LORTOT
INSTRUCTOR	HASUBT2	0.9561 0.0001	0.5176 0.0002	0.3073 0.0378	0.7407 0.0001
	OASUBT2	0.5574 0.0001	0.9400 0.0001	0.6397 0.0001	0.8606 0.0001
	SASUBT2	0.4193 0.0037	0.6089 0.0001	0.9457 0.0001	0.7012 0.0001
	LORTOT2	0.7923 0.0001	0.8636 0.0001	0.7038 0.0001	0.9486 0.0001

Hasubt =LORS Subscale of Home Activities
Oasubt =LORS Subscale of Outside Activities
Sasubt =LORS Subscale of Social Activities
Lortot =LORS Total Score

Part II: Validity Study

Establishing the Validity of the Study's Functional Scales

In examining the validity of the study scales, the Functional Status Assessment Instrument (FSAI) designed by Jette (1978) served as the criterion for comparison. The FSAI was matched to like-items of Self Care and Mobility from the Barthel Index while the LORS' items of Home Activities, Outside and Social Interaction were compared to similar entries in the Jette Scale. From these indices, eight separate divisions of functional status were organized; four for the Barthel-Jette comparison and four for the LORS-Jette match. This data was then examined for the presence of statistical associations between the individual item comparisons using Spearman Correlation procedures. The FSAI was scored on a decreasing scale while both the Barthel Index and the LORS were measured on an ascending scale, therefore, an inverse relationship was anticipated.

The first comparison examined the Self Care component (Table 4-20). Three items from the Barthel Scale were compared to four similar items from the Functional Status Assessment Index. The direction and magnitude of the relationship ranged from $r=-0.71$, ($p<0.0001$) for the items of "Upper Body", (Barthel Scale) and "Pants" (FSAI) to $r=-0.93$, ($p<0.0001$) for the items of "Lower Body", (Barthel Scale) and "Shoe", (FSAI).

The subscale of Mobility was then constrasted for like

items of "Walk", "Wash", "Toilet" and "Stairs" (Table 4-21). Here, statistically significant negative correlations were achieved ranging from $r=-0.52$, ($p<0.0002$) for the variable "Walk" to $r=-0.89$, ($p<0.0001$) for the use of stairs.

Various transfer activities were then compared (Table 4-22). Here again, fair to good associations were seen in the 3x3 matrix. Similar functions of washing and getting into the tub or shower actually rated lower than expected $r=-0.59$, ($p<0.0001$) while good relatedness was recorded for toilet and bed transfers with values ranging from $r=-0.74$, ($p<0.0001$) to $r=-0.77$, ($p<0.0001$).

Hand Activities made up the fourth component of comparison which was built of items from the Barthel, LORS and the Jette Scales (Table 4-23). The Barthel variables of "Cup", and "Eat" and the LORS' item of "Simple Foods" were matched with the variables of "Cutfood", "Writing", "Open Container" and "Faucet" from the FSAI. Once again, significant negative correlations were obtained. Interestingly, what appeared in theory to be a close match of items, like "Simple foods" (LORS) and "Cut food" (FSAI) resulted in the lowest correlation for this group ($r=-0.39$, $p<0.008$). On the other hand, excellent association was evident for the Barthel item "Eat" and the FSAI item "Cutfood". Clearly, all variables in this subdivision of Hand Activities showed some relationship, most were related to food whether in the preparation or eating. The one exception was the item of "Writing" which demonstrated fairly low yet significant correlations. However, given that this term

lacked any direct relation to food preparation or consumption, the correlations of $r=-0.45$, ($p<0.001$) to $r=-0.58$, ($p<0.0001$) were more than satisfactory.

Table 4-20

Spearman Correlations for the Variables of the Barthel Index and the Functional Status Assessment Instrument for the Status of Self Care

SELF CARE

(N=46)

BARTHEL INDEX

		UPBODY	LOWBODY	GROOM
FSAI	BUTTON	-0.92209*	-0.85916*	-0.81482*
	PANTS	-0.70959*	-0.77907*	-0.77749*
	SHOES	-0.84662*	-0.93046*	-0.91936*
	HAIR	-0.84128*	-0.78528*	-0.90014*

* $p<0.0001$

Barthel Index Upbody =dressing upperbody
Barthel Index Lowbody=dressing lowerbody
Barthel Index Groom =comb hair and brush teeth
FSAI Button=fasten buttons
FSAI Pants =put on pants
FSAI Shoes =put on shoes
FSAI Hair =comb hair

TABLE 4-21

Spearman Correlations for the Variables of the Barthel Index and the Functional Status Assessment Instrument for the Status of Mobility

MOBILITY

(N=46)

		BARTHEL INDEX			
		WALK1	WASH1	TRATOIL1	STAIR1
FSAI	WALK2	-0.5238 0.0002	-0.5768 0.0001	-0.6314 0.0001	-0.5488 0.0001
	WASH2	-0.2283 0.1314	-0.6011 0.0001	-0.5257 0.0002	-0.3887 0.0083
	TRATOIL2	-0.5802 0.0001	-0.5755 0.0001	-0.7371 0.0001	-0.5024 0.0004
	STAIR2	-0.7563 0.0001	-0.4524 0.0016	-0.5480 0.0001	-0.8892 0.0001

Walk1-Walk2=the Barthel variable "Walk" compared to the "Walk" variable of the FSAI

Wash1-Wash2=the Barthel variable "Tub-Shower" compared to the "Wash" variable of the FSAI

Tratoil1-Tratoil2=the Barthel variable "Transfer to Toilet" compared to the same variable of the FSAI

Stair1-Stair2=the Barthel variable "Stairs" compared to "Stair Climbing" of the FSAI

TABLE 4-22

Spearman Correlations for the Variables of the Barthel Index and the Functional Status Assessment Instrument for Transfer Ability Status

TRANSMOBILITY

(N=46)

BARTHEL INDEX

		TRACHAIR	TUBSHOW	TRATOIL1
FSAI	TRABED	-0.7678 0.0001	-0.3951 0.0066	-0.7678 0.0001
	WASH2	-0.5257 0.0002	-0.5867 0.0001	-0.5257 0.0002
	TRATOIL2	-0.7371 0.0001	-0.4316 0.0027	-0.7371 0.0001

Trachair =Barthel variable "Transfer to a Chair"
Tubshow =Barthel variable "Transfer to Tub or Shower"
Tratoil1 =Barthel variable "Transfer to the Toilet"
Trabed =FSAI variable "Transfer to bed"
Wash2 =FSAI variable "Maneuvring about the sink or tub"
Tratoil2 =FSAI variable "Transfer to the Toilet"

TABLE 4-23

Spearman Correlations for the Variables of the Barthel Index, the LORS and the Functional Status Assessment Instrument for the Hand Activities Status

HAND ACTIVITIES

(N=46)

BARTHEL INDEX and LORS

		CUP	EAT	SIMFOOD
FSAI	CUTFOOD	-0.7067 0.0001	-0.9979 0.0001	-0.3963 0.0085
	WRITING	-0.4513 0.0016	-0.4513 0.0016	-0.5821 0.0001
	OPENCONT	-0.4572 0.0014	-0.4207 0.0036	-0.5887 0.0001
	FAUCET	-0.6042 0.0001	-0.6019 0.0001	-0.7297 0.0001

Cup =Barthel variable "Drinks from Cup"
Eat =Barthel variable "Eating"
Simfood =LORS variable "Preparing Simple Foods"
Cutfood =FSAI variable "Cutting food"
Writing =FSAI variable "Writing"
Opencont=FSAI variable "Opening a Container"
Faucet =FSAI variable "Turning on a Faucet"

The Level of Rehabilitation Scale was then compared to equilivant variables from the Functional Status Assessment Instrument. The first testing examined specific activities performed in the home (Table 4-24). Fifteen of the thirty-five items collated showed varing degrees of association. Of particular interest was the matching of the LORS variable of "Lightwork" to seven work-related items of the FSAI. Six of the seven comparisons demonstrated statistically correlations with values ranging from $r=-0.33$, ($p<0.02$) for the "Lightwork-Yardwork" match to $r=-0.67$, ($p<0.0001$) for the "Job Responsibilities-Lightwork" combination. However, there was no significant correlation exhibited for the FSAI variable "Washing Windows" and the LORS variable of "Lightwork". Similarily, the variable of "Heavywork" did not relate with specific home tasks from the FSAI. Surprisingly, there was no correlation between the variables of "Landry" and "Heavywork". On the other hand, the LORS variable of "Odd Jobs" demonstrated significant associations with the FSAI items of "Writing", $r=-0.50$, ($p<0.0005$) and "Job Responsibilities", $r=-0.41$, ($p<0.004$). Furthermore, the LORS item of "Pastime" was significantly related to the FSAI variables of "Vacuum", $r=-0.34$ ($p<0.02$), "Cupboard", $r=-0.35$, ($p<0.01$), "Writing", $r=-0.36$ ($p<0.01$) and "Job Responsibilities", $r=-0.34$ ($p<0.02$).

Outside Activities were next compared from both the LORS and the FSAI (Table 4-25). Here, over 70% of the items showed some degree of significant association. The FSAI item of walking proved to be statistically related to "Outside

Activities, r=-0.46 ($p<0.005$), "Shopping-Errands", r=-0.40 ($p<0.006$), "Taking transportation independently", r=-0.32 ($p<0.04$) and "Independence in taking long trips", r=-0.70 ($p<0.003$). Likewise, the FSAI variable of "Job Responsibilities" showed fair correlation with the LORS items of "Outside Activities", r=-0.34 ($p<0.02$); "Shopping-Errands", r=-0.44 ($p<0.002$); "Spectator Events", r=-0.33, ($p<0.04$); "Transportation Independently", r=-0.50 ($p<0.001$); and "Long Trips Accompanied", r=-0.44 ($p<0.002$).

Social Activities was the last section to be examined (Table 4-26). Again, statistically significant associations were present between the FSAI variable of Socialization and the three LORS items of Home, Outside and Church-Synagogue Socialization with values ranging from r=-0.40 ($p<0.01$) to r=-0.57 ($p<0.0001$).

In sum, excellent correlations were obtained for the individual item comparisons of Self Care, Mobility, and Transfers Components from both the Barthel Index and the Functional Status Assessment Index. In addition, similar hand activities from the Barthel Index, the LORS, and the FSAI produced fair to good correlations for the variables selected. When equitable variables from the Level of Rehabilitation Scale and the Functional Status Assessment Index were compared, 43% of the Home Activities items, 71% of the Outside Activities, and 50% of the items related to Social Interaction produced statistically significant associations.

TABLE 4-24

Spearman Correlations for the Variables of the Level of Rehabilitation Scale and the Functional Status Assessment Instrument For the Home Activities Status

		HOME ACTIVITIES (N=46) LORS				
		LIGHTWOR	HEAVYWOR	ODDJOB	PASTIME	TETEL
FSAI	VACUUM	-0.4964 0.0007	-0.1987 0.2013	-0.2631 0.0808	-0.3363 0.0223	-0.1623 0.2868
	CUPBOARD	-0.3878 0.0102	-0.2523 0.1025	-0.3245 0.0296	-0.3516 0.0166	-0.6232 0.0001
	LANDRY	-0.4852 0.0010	0.0000 1.0000	-0.2624 0.0816	-0.1208 0.4239	-0.0084 0.9559
	WASHWIND	-0.2614 0.0904	-0.0673 0.6680	-0.0675 0.6594	-0.0771 0.6105	-0.2277 0.1325
	YARDWORK	-0.3448 0.0235	0.0835 0.5943	-0.1490 0.3286	-0.1487 0.3238	-0.1535 0.3140
	WRITING	0.5250 0.0003	-0.0925 0.5550	-0.4970 0.0005	-0.3642 0.0128	-0.4615 0.0014
	JOBRESP	-0.6769 0.0001	-0.2427 0.1168	-0.4138 0.0047	-0.3416 0.0201	-0.2334 0.1228

CODE

Lightwor=Lightwork
Heavywor=Heavywork
Oddjob =Odd Jobs
TelTel =Telephone-Television
Washwind=Wash Windows
Jobresp =Job Responsibilities

TABLE 4-25

Spearman Correlations for the Variables of the Level of Rehabilitation Scale and the Functional Status Assessment Instrument For the Outside Activities Status

		OUTSIDE ACTIVITIES						
		LORS						
		OUTACTIV	SHOPERR	SPECTEV	TRANSPAC	TRANSPIN	LONGTRIA	LONGTRII
FSAI	WALK2	-0.4558 0.0015 46	-0.3951 0.0066 46	-0.1467 0.3861 37	-0.1045 0.4891 46	-0.3271 0.0420 37	-0.1601 0.2876 46	-0.7014 0.0036 15
	JOBRESP	-0.3408 0.0205 46	-0.4357 0.0025 46	-0.3350 0.0426 37	-0.4766 0.0008 46	-0.49510 0.0014 37	-0.44163 0.0021 46	-0.34915 0.2021 15

Code

Outactiv=Outside Activities
Shoperr =Shopping and Errands
Spectev =Spectator Events
Transpac=Transportation Accompanied
Transpin=Transportation Independently
Longtria=Long Trips Accompanied
Longtrii=Long Trips Independently
Walk2 =FSAI variable of Walk
Jobresp =Job Responsiblities

TABLE 4-26

Spearman Correlations for the Variables of the Level of Rehabilitation Scale and the Functional Status Assessment Instrument for the Social Activities Status

SOCIAL ACTIVITIES

LORS

		HOMESOC	OUTSOC	CHURSYNA
FSAI	SOCIALIZ	-0.5661 0.0001 46	-0.3069 0.0403 45	-0.3953 0.0170 36
	CHURCH	-0.1852 0.2178 46	-0.1675 0.2714 45	-0.0725 0.6741 36

Homesoc =LORS variable of "Home Socialization"
Outsoc =LORS variable of "Outside Socialization"
Chursyna=LORS variable of "Church or Synagogue"
Socializ=FSAI variable of "Socialization"
Church =FSAI variable of "Going to Church"

CHAPTER V

Discussion

The Results Chapter presented two separate investigations, the examination of the rater reliability and the validation of a pair of functional status indices used in the Geriatric Trial. To estimate the reliability of the patient assessors, the overall variation among the 23 raters and the study norm was compared when the Barthel Index and Level of Rehabilitation Scale were used to estimate patients' status. Consideration was also given to the inter-observer and the intra-observer reliability of all study participants. The primary testing of rater reliability was achieved through the use of videotaped interviews of six patients evaluations. These tapes were presented to all raters in a test-retest sequence. In an effort to evaluate adherence to study protocol and the continued reproducibility of rater scores, on-the-spot inspections were conducted over a one year period in both the hospital and home settings.

The Barthel Index and LORS instruments were tested for Concurrent Validity (Criterion-Related) using the Functional Status Assessment Instrument (FSAI) designed by Jette (1980). The three scales were examined by means of individual item comparisons and levels of association were tested for significance.

The study raters were classified into three separate

groups. The first group consisted of 14 evaluators who were responsible for the data collection for the Geriatric Trial. Five interpreters were employed to help the research assistant and the study evaluators in communication and data collection procedures with patients who spoke neither French nor English. They made up the second group of raters. The third group was made up of four instructors responsible for rater training and the overall organization of the Geriatric Trial. In this discussion, each of these sections on reliability and validity will be considered separately.

Overall Variation between the Raters and the Gold Standard

As demonstrated in Chapter IV (Results), the data generated from the measurement indices contained varying degrees of measurement error. The fact that variation existed was not the primary concern of this study. Rather it was the sources of variability that were the important elements in determining the rater reliability of the study scales.

For this reason, several methods were employed to assess the extent of the rater variability when the Barthel Index and the Level of Rehabilitation Scale were used as assessment tools. Descriptive measures were first obtained by examining the mean scores and the standard deviations for all groups and compared to the Standard. The Standard Deviation (SD) was used as a measure of variation instead of the Standard Error (SE) because the objective of this study was to determine the magnitude of the variation present. As the Standard Error

will always be smaller than the Standard Deviation, it was felt that the Standard Error would be inappropriate in this situation because it would not adequately describe the spectrum of the data collected. Feinstein (1985) reported that the standard error was often used improperly for descriptive purposes, resulting in a distorted image of the data because the SE denoted the fragility of the mean and not the scope of the variation for the data in question.

Agreement ratios were then calculated to establish the percentage of concordance between the twenty-three raters and the study's Gold Standard. In general, all participants demonstrated fair to excellent agreements with the Standard for the first two subscales of the Barthel Index. However, there was a marked decrease in agreement for the subsection "Mobility" and for "Total Status". A suggested reason for this discrepancy could be that mobility was defined as the ability to walk 50 yards or to negotiate a wheelchair independently. This seemed ambiguous to the Evaluators. For example, "patient two", had suffered a stroke with resultant hemiplegia. Although the patient felt that she was capable of walking the required distance, several raters judged her to be dependent because of the effort that was required for her to accomplish this task. In constrast, "patient three" had amputations of both lower extremities and therefore was confined to a wheelchair. Nevertheless, he demonstrated a definite freedom of mobility in his chair, leaving little doubt of his personal independence.

The Total Status variable of the Barthel Index was a summary score describing the overall functional status of the individual. Therefore, it was not surprising to see relatively low agreement ratios, as a masking effect was present when the scores of the three subsections of the Barthel Index were combined.

Poor to fair rater agreement ratios were seen for all subsections of the Level of Rehabilitation Scale. Perhaps, in constrast to the Barthel Index, the LORS achieved lower ratios because it addressed broader issues of home and community independence. Specificity of tasks within each subscale was not explicit, rather the rater was left to decide from several examples the activities of the patient that best described the level of functioning. As a result, the chances of one to one agreement was proportionally decreased as the activities within the subscale became more diversified.

From these agreement ratios, estimates of rater measurement bias with the Gold Standard were computed for the two study scales. In general, the raters tended to underestimate the functional status of the six patients particularly when using the Level of Rehabilitation Scale. This tendency may be related to the clinical background of the Evaluators. From a practical standpoint the primary concern in evaluating the functional status of a patient is to estimate the patient's level of dependence and the level of assistance required to permit that person to live as independently as possible. Clinically speaking, it is

therefore better to slightly underestimate than over inflate the patients' level of function. This will ensure that a careful screening of needs occurs and that an adequate support network will be provided to allow the individual to reassume her or his role in the community. It appears that the clinical background of the Evaluators, ingrained overtime, may remain preeminent in spite of the specific teaching and practice sessions given for the purposes of training study assessors. While the outlook is no doubt beneficial to the patient, from the study's point of view, it is possible that this systematic error could lead to false conclusions.

In sum, this descriptive data provided an overall impression of the percentage of variation that existed between the raters and the Gold Standard for the two study scales. The next step was the identification of sources and the extent of variability among the rater groups.

Through the analysis of variance, the intraclass correlation coefficient "R" was computed to estimate the stability of each group's position with the Gold Standard. The intraclass coefficient is an "index of concordance for continuous data which combines a measure of correlation with a test of the difference in means" (Bartko, 1966). This index assesses not only the similarity of slopes, but also the similarity of intercepts. Therefore, if one individual is systematically higher or lower than the other, the intraclass correlation coefficient (ICC) will reflect this bias (Kramer and Feinstein, 1981). The ICC is derived from a repeated

measures analysis of variance.

Using this procedure, the Barthel Index was the first to be tested for its level of rater reliability. The findings were generally in disagreement with most studies reported in the literature. A possible explanation for this difference was that variations were minimal among the patients selected for the videotaping in two subscales of the Barthel Index. All six patients were relatively equivalent in their levels of independence for self care and continence. These similarities among patients virtually eliminated the primary source of variability normally expected in this type of analysis, that is the variation due to patient biological differences. Because patient variance is a characteristic of the population studied, whereas the variance of random error is essentially a function of the measurement procedure, reliable measurements are easier in a heterogenous population than in a homogeneous population (Fleiss et al., 1977). As a result, the variation due to specific raters and patients took precedent.

Although at the time of the videotaping, the participating patients appeared to have different levels of functional status, in retrospect, it should have been anticipated that similarities existed. One reason for the homogeneity in self care activities and continence was related to the fact that these patients were attending a Day Center at a rehabilitation institute. This implied that the person had most likely achieved a level of independence necessary to function in the community. The two basic functions generally required are the management of self care and continence

activities.

This lack of patient variation specifically affected the overall agreement with the Gold Standard for the Evaluators and the Interpreters. It was reassuring, however, that the Instructors achieved perfect agreement despite this lack of patient variability since one of the major incentives to test the group of Instructors was to examine the presence of systematic bias between the study organizers themselves.

The agreement levels for the remaining two subscales of the Barthel Index reflected acceptable levels of concordance for each of the groups of raters with the Gold Standard. At first, there appeared to be a poor level of agreement among the 23 raters and the Standard for the variables "Mobility" and "Total Status". However, in the initial examination of the data, the main concern was to determine the percentage of rater agreement for each of the six patients. This information provided insight into those raters who demonstrated the greatest indecision in recording functional status when the patients presented ambiguous levels of function. The data collectors were then evaluated according to their group assignment. As seen from the coefficients for the "Self Care" and "Continence" subscales, the Instructors continued to establish excellent levels of agreement in the areas of "Mobility" and "Total Status". Similar coefficients of concordance were seen for the Interpreters while the Evaluators were registered as having only fair agreement for the same subsections although stochastic significance was obtained. It has been reported, however, that quantitative

significance of the intraclass correlation depends on its absolute magnitude (Kramer and Feinstein, 1981). Burdock and colleagues (1963) suggested an ICC=+0.75 as being accepted. This value implies that there would be little residual variablility present to confound good discrimations among the subjects.

Since the coefficients generated by the Evaluators for these two subscales were well below the recommended level of R=0.75, these results insinuate that although rater consistency was present, conformity with the Standard was not. Accordingly, specific tests of conformity were carried out to estimate the levels of significance. Clearly, systematic bias existed between the Evaluators and the Standard for the variable Total Status whereas the Mobility subscale while not statistically significant, was borderline.

On average, the Barthel Scale appeared to be well understood by the Interpreters and the Instructors as reflected in the Total Status scores (Interpreters: $R=0.79$, $p<0.0001$; Instructors: $R=0.88$, $p<0.0001$), but the Evaluators continued to demonstrate relatively low, yet significant, agreements ($R=0.55$, $p<0.0001$) for this measure. This finding suggests that the rater-training program was insufficient for the diverse types of raters employed in the study. The group of Evaluators were predominantly from a health care background, whereas, the Interpreters were translators of a particular language and had no specific health related education. The initial training program was designed to introduce the field of Geriatrics to all those unfamiliar with

the area. As a result, the major emphasis was directed towards those raters with minimal prior knowledge in the hope of minimizing their potential inconsistencies. What was neglected, however, was the fact that other forms of group bias are frequently present. For example, significant biases may be present among the professional groups themselves. In retrospect, this would seem to have been the case in this particular investigation.

Unlike the Barthel Index, agreement ratios for the Level of Rehabilitation Scale were well within the acceptable levels in establishing consistency for all three groups of raters. However, when the tests of conformity were calculated, statistically significant bias was evident for the Interpreters in the area of Home Activities whereas both the Evaluators and Interpreters were systemically lower than the Standard for the variable Total Status. This rater bias although statistically significant was marginal from a clinical point of view. This results, nevertheless, signaled potential areas of future ambiguity among the raters when assessing patients by means of the LORS.

In sum, there appeared to be better reliability between the raters and the Gold Standard for the Level of Rehabilitation Scale than for the Barthel Index. Clearly, rater variation was greatest for the subscales Self Care and Continence. The main reasons for this marked discrepancy can be attributed to three factors: certain interviewers reported difficulties in interpreting the guidelines in the instruction

manual for the domain of continence; a few raters tended to define the patients' functional status inappropriately; and the six patients chosen for the testing of rater reliability were remarkedly similar in their levels of Self Care and Continence. As a result, the variance due to differences between raters and patients was greater than the variance apportioned to the patient themselves.

Inter-rater Reliability

In assessing the between-rater reliability, a familiar pattern of results became evident. Once again, the mean scores and the standard deviations of rater recordings provided the first insight into the levels of agreement between the three groups of raters. Although group scores ranged within the predetermined acceptable spread, the Evaluators continued to be the most conservative of the raters in scoring the six patients. This trend was relatively consistent for both the Barthel Index and the LORS.

The objective of the analysis was to determine the degree of inter-rater variation for each functional scale. In recent years, measurement theorists have attempted to promote the use of appropriate procedures to assess the consistency or the reliability of a tool. Traditionally, the Product Moment Correlation approach has been used most frequently in determining the levels of concordance between raters administering the same instrument. This technique, however, has one serious drawback often ignored by those who choose to

use it. Although an association may be obtained between different raters there is no way of estimating if systematic bias exists between these same raters. Therefore, a test could appear "reliable" yet may not be "valid".

A more appropriate approach to the measurement of concordance is the Intraclass Correlation Coefficient "R". At this point, note should be made as to why both procedures were employed in this evaluation of rater reliability. The objective was to delineate these approaches in order to clearly understand the advantages and disadvantages of each method. The results attained, however, were unexpected. Although the Pearson Correlation generally followed a similar pattern to the Intraclass Coefficients for both of the scales, the values were distinctly lower, rather than higher, as expected. Product Moment Correlations for the Mobility subscale and the Total Status Variable of Barthel Index, in particular, were considerably less than the Intraclass Coefficients for the same items. One possible explanation for this situation could be that because the Product Moment Correlation is a bivariate statistic, it is meant to be used in the determination of the relationship between two variables. However, when two different raters or groups of raters assess the same task only one variable is being measured. Because the data must be reduced to two sets of scores for correlation, tests that involve multiple raters must be divided and averaged with the resultant loss of information. In essence, the Product Moment Correlation is unable to take different sources of variance into account. In

consequence it is clearly inappropriate as a measure of consistency among several raters (Hasselkus, 1976).

The Analysis of Variance, on the other hand, is a statistical procedure that does exactly what its name implies, that is, it analyzes the variance of the data. Through the identification of the sources of variation, the first steps towards estimating reliability can be achieved. For this study, reliability of the raters was expressed by the Intraclass Correlation which corresponded to the proportion of the true variance among the patients being measured divided by the total variance of the data. As there is no loss of information with this procedure, the estimates of inter-rater reliability calculated by this method were definitely stronger than those produced by the simple correlation procedures for both the Barthel Index and the LORS. As discussed earlier, the ICC is subject to one problem when there is a lack of true patient variation. In these situations, the total variance is greater than the true variance which results in a spurious estimate of reliability. Such was the case for the Barthel subscales of Self Care and Continence. Although little inference could be made from these estimates of reliability, the delineation of the sources of variation through the Analysis of Variance, permitted a clearer understanding of the percentage of variation attributed to the remaining elements of the total variance for these subscales. In both cases, the greatest source of variation was assigned to differences between specific raters and patients.

The other major advantage of using the Intraclass

Correlation as opposed to the Product Moment Correlation is that it provides a means of determining the presences of systematic bias between raters. This bias was visible for the each group comparisons for the Barthel variables, of "Mobility" and "Total Status". Furthermore, although the coefficient of reliability was low (R=0.17) between the Evaluators and the Instructors, for the Self Care subscale, systematic bias could still be identified between these groups. In constrast, there was a total absence of systematic variation between the groups of raters for the Level of Rehabilitation Scale.

These results were unexpected. Initially, the investigators had felt that the LORS might be a potential source of indecision among raters because of the subjectivity that could be built into the different levels of the scale. Further, this index which was designed for a younger population placed emphasis on certain tasks that generally were not considered as being representative of the older age group. It was, therefore, somewhat of a surprise to see a strong degree of association between the three groups of raters when using this scale. In contrast, the Barthel Scale, as reported in the literature, was reported to have good psychometric properties (Granger, 1979). The index has demonstrated a test-retest reliability of 0.89 and an inter-rater reliability of 0.95 (Granger, 1979). Additionally, Sherwood and colleagues (1977) have described high alpha reliabilities ranging from 0.95 to 0.96 thus suggesting that the test was internally consistent as a

measure of self-care abilities. Thus, it was equally surprising to observe the scope of correlations ranging from 0.00 to 0.90 for this study. One suggested explanation for these results was attributed to the fact that there was relatively little variation among the raters themselves for the two subscales, Self Care and Continence. When the agreement ratios were re-examined, it was the variable of Mobility that was the principle source of variation in the Barthel Index. Therefore, when ICC procedures were performed for the subscale of Self Care and Continence, the deviations about the mean were almost entirely non-existent, thus producing Intraclass Correlations with questionable findings.

A second explanation for this poor showing of the Barthel Index may stem from the fact that all reports of this scale have been generated from evaluations made by care givers (Granger, 1979; Gresham, 1980; Sherwood, 1977; Wylie, 1967) and not by independent assessors employed in a clinical trial. Consequently, the high levels of reliability for the Barthel Index would seem to be reflecting mainly clinical judgements formulated overtime rather than independent and "one shot" assessments of physical status.

The LORS instrument, on the other hand, produced scores with more uniform degrees of variation. As a result, the associations of the raters scores could be examined. The reasons for this discrepancy have already been proposed in the rater-gold standard comparison.

In brief, the Intraclass Correlation Coefficient permitted a more detailed picture of the components of the

equation which contributed or detracted from the overall estimate of reliability between the different raters.

Intra-rater Reliability

The rater scores for both Indices were then examined over time using the Product Moment and the Intraclass Correlation procedures. In this comparison, outstanding differences in coefficients were seen for the subscales Self Care and Continence of the Barthel Index. In particular, there was a distinct discrepancy in the two approaches when the Interpreter group was assessed. Once again, these results were severely affected by the lack of variation among the six patients for the areas of Self Care and Continence, yet the ICC still was able to reflect a change in the Interpreter functional scores over time. What appeared to be low within-rater estimates for the Evaluators in the area of Continence through the Product Moment Correlation proved to be zero agreement with the Intraclass procedure. These measures indicated that poor stability existed for these subscales for the first two groups of raters. A lack of sufficient understanding or individual interpretation of the instructive guidelines for these areas could explain these findings.

A total contrast was seen in the Self Care and Continence correlations for the Instructor group. What was recorded as zero agreement for the simple correlation procedure was listed as complete agreement through the ICC. Although patient variation was minimal, as stated previously,

when the raw data was reexamined it was clear that no change had occurred in the Instructor scores from time one to time two. However, the Product Moment Correlation was unable to pick up this total lack of variation for these scores. In other words, the deviations about the mean were non-existent thus producing a zero estimate. For the remaining sections of the Barthel Index and the LORS, relatively similar correlations were obtained from both procedures which indicated that fair to excellent intra-rater reliability had been present for all raters.

As reported in the literature (Hasselkus, 1976; Kramer and Feinstein, 1981), the Product Moment Correlation Coefficient proved to be an unsatisfactory statistical procedure for this study in terms of estimating between and with-in rater agreement. The Intraclass Correlation Coefficient, on the other hand, provided a measure of intrinsic accuracy of the two study instruments.

Program Re-evaluation

The evaluation of rater reliability was initiated early in the Clinical Trial to determine the quality of the data collected by the raters employed by the Geriatric Study. From these results, it was evident that the definitions and guidelines describing the variables Self Care, Continence and Mobility needed to be restated. The problem of the unacceptable rater variation for the Barthel Index also had to be addressed. Therefore, the first procedure involved the

editing of the users' instruction manual to improve the overall clarity of the terms of reference and to reduce rater-expressed ambiguities. Additional patient evaluations were then scheduled for those raters who had demonstrated the greatest discrepancies. Each individual conducted three, separate, randomly selected interviews and simultaneous scoring was recorded by both the rater and one of the study instructors. Pair-wise agreement ratios were once again calculated using the Intraclass Correlation procedure. The results of these tests indicated that the overall agreement had markedly improved with estimates ranging from R=0.75 to 1.00 ($p<0.0001$) for both scales. As reported by Fleiss (1977), in obtaining replicate ratings on each subject and averaging them, the errors of measurement averaged out and the reliability coefficients increased.

On-The-Spot Inspections and Adherence to Study Protocol

Early assessment of acceptable rater reliability, however, does not guarantee continued high levels of agreement over the length of a two year Trial. In order to ensure that rater reliability continued to reach acceptable levels and to monitor adherence to study protocol, on-the-spot inspections were conducted for the life of the Geriatric Trial in both the hospital and home settings. Through this monitoring of raters, excellent agreement levels continued to be achieved for both the Barthel Index and the Level of Rehabilitation Scale. Rater comprehension for the functional activities of

Self Care, Continence and Mobility was no longer a problem.

In a study like the Geriatric Trial, results can be controversial or can influence future policy decisions, therefore, it becomes increasingly important that adequate quality control procedures are conducted. The issues of systematic bias are constantly an area of concern, and need to be reassessed over the designated life of a trial. Likewise, the question of blind assessments has to addressed in order to insure that any differences seen between the two treatement procedures are in fact true differences and not simply rater partiality.

In summary, the overall rater reliability as well as the inter-rater and intra-rater concordance fell within acceptable limits for the Level of Rehabilitation Scale. The scoring of the Barthel Index, on the other hand, did not achieve good rater agreement in the initial testing sessions. After thorough examination of the results, the instruction manual was rewritten in excruciating detail to clarify the Barthel Index specifications. Individual raters were retrained and retested. Subsequentally, excellent agreement ratios were obtained and continued to be present over the duration of the Geriatric Study. Although the LORS instrument was the source of greatest variability, this variation was predominately attributed to the patients themselves. As a result, good to excellent agreements were achieved for this scale. Furthermore, there was no evidence to indicate that

significant systematic bias was present between the raters when this scale was used. This was not the case for the Rater-Standard comparison. Nevertheless, only limited bias was present for this comparison and this was related specifically to differences between the Interpreters and the Gold Standard. Although the twenty-three raters tended to underestimate the status of the video-taped patients, this bias was reduced to near zero after additional instructions and periodic inspections were introduced into the program. Individual rater differences continued to be present, yet, the study organizers felt that overall rater agreement had been attained within acceptable limits. In sum, by establishing a satisfactory level of rater reliability for the two study scales, full emphasis could then be directed to the estimation of the differences in the outcome measures for the Parent Investigation of Geriatric Care.

As Knatterud (1979) stated, an error-free study may be laudable but not feasible nor practical. In brief, an error-free study is not a reasonable goal. What is important, however, is that the number of errors remains small and that the errors present are randomly distributed among the participating groups. But above all, these goals should be obtained at a reasonable price in terms of the development of a quality assessment program and the assurance of quality data handling and processing.

Validity Study

Validation of Study Instruments

The two functional instruments, the Barthel Index and the Level of Rehabilitation Scale (LORS) were compared to the Functional Status Assessment Instrument (FSAI) through the correlation of items held to be valid measures of the same domain. This is defined as criterion-related validity, a concurrent approach. As the scales had different scoring systems, item by item comparisons were made rather than the collating of composite scores. Eight separate divisions of functional status were created which included Self Care, Mobility, Transfers, Hand, Home, and Outside Activities, as well as, Social Interaction. The first four divisions represented the Barthel-FSAI comparison while the last four divisions made up the LORS-FSAI comparison.

For the area of Self Care, consistently good to excellent negative relationships were present between similar items from the Barthel-FSAI. These associations were reassuring because they seemed to indicate that those items which referred to independence in personal care possessed comparable meanings, thus, measuring the domain they claimed to measure, that being the concept of Self Care ability. Correlations for the dimensions of Mobility were also statistically significant for the like-variables of "Walk", "Wash", "Toilet" and "Stairs". Again, these results supported the belief that the activities of ambulation had been addressed. It would appear that

slightly different meanings made up the dimensions of "Walk" and "Wash", as these correlations ranged between $r=-0.52$, ($p<0.0002$) and $r=-0.60$, ($p<0.0001$). Nevertheless, they were acceptable comparisons. The variables labelled "Transfer to toilet" and the "Use of stairs", on the other hand, were much more precise which appeared to be mirrored in the resultant coefficients. In the next comparison, a slight decrease in the strength of associations was seen for Transfer Activities, even though statistical significance was again achieved. Interestingly, as the functions became more complex, leaving more room for interpretation, the corresponding correlations seemed to reflect the variations within the specific activities. The next section addressed the use of the hands, in particular. Again, fair to excellent relationships could be interpreted from these correlations. It was hypothesized that these matched items represented related hand functions only, therefore, strong associations had not been expected for this domain. These results, however, might be explained by the fact that similar anatomical movements of the hand, although not performing the exact same function, frequently require the same dexterity. For example, if an individual prepares simple foods, it would be conceivable that a container may need to be opened or a faucet might be turned on during the activity of food preparation which could explain the correlations seen.

In essences, the comparison of analogous items from both the Barthel Index and the Functional Status Assessment Instrument provided supportable evidence that indeed the

domain of personal physical function was being measured as was hypothesized. This proved to be an important outcome in the validation of the Barthel Index for this geriatric popualtion.

The pairing of items from the LORS and the FSAI were not as impressive as those just seen for areas of physical status. Within the scope of Home Activities, fair to moderate correlations were present for the related items of "lightwork", "vacuuming", and "laundry" yet no associations were seen when the area of "heavywork" was addressed. This was surprising because one could argue that "washing windows" and "yardwork" is harder and requires more energy than simpler activities within the house. On the other hand, it is possible to encounter what appears to be, as Colton (1974) pointed out, nosense or spurious correlations between two variables that logically seem to be unrelated to one another. This could have been the situation for the variables of "Cupboard" and "Odd Jobs" (-0.32, $p<0.02$) as well as the "Telephone-Television- Cupboard" comparison (-0.62, $p<0.0001$) and the "TelTel-Writing" combination (-0.46, $p<0.001$). No further explanations could be found for these results. Nevertheless, the associations that were present supported the claim that acceptable degrees of validity were evident for this portion of the total scale.

When the domain of Outdoor Activities was assessed, once again, a fair relationship was seen for better than 70% of the items compared. These findings were reassuring even though the correlations were relatively low, since the aim was to establish some trend between these pooled variables. Clearly,

the definition of "Outside Activities" can have a broad meaning. For our purposes, therefore, the items of the LORS and the FSAI appeared to relate to similar domains.

The last division of functional status that needed to be assessed was that of "Social Interaction". Here, fair yet statistically significant relations were seen for the various elements of socialization reported in the two scales. What was surprising, however, was the fact that the area of socialization could include going to or participating in activities in a church of synagogue but the same association was not seen in the reverse. In other words, the item of "Church" did not imply "Socialization".

In summary, the Barthel Index and the LORS were tested for validity by comparing the items of these scales to a third instrument, the Functional Status Assessment Index (FSAI). Better correlation coefficients were seen for the Barthel-FSAI comparisons. This was satisfying but not surprising as the two scales used similar variables for the areas of "Self Care" and "Mobility". For the LORS-FSAI comparisons, fair to good correlations were seen for several items yet there were several items that showed no association. Surprisingly, there was essentially no relationship between the LORS item of Church-Synagogue and the FSAI variable of Church. One possible explanation of this could be that this item was rarely scored when the elderly person was interviewed. Better than 90% of the responses were listed as non-applicable. As a result, it was evident that this item did not represent the

population understudy.

To draw inference from these findings, it could be said that more than three quarters of the items compared from the three study scales proved to have at least fair to moderate relationships. In some instances, particularly those activities that were related to Self Care, Mobility, Transfer Activities and Hand Function consistently good correlations were encountered. On the other hand, it was clear that the indices selected for the Geriatric Trial were not always designed specifically for the elderly clientele as certain items such as Church, Spectator Events, Heavywork, Telephone-Television were frequently no longer an interest or a necessity for the older person. Many of the visited subjects who were still in the community lived with someone younger and therefore these tasks were assumed by that person. Nevertheless, given these limitations, over 75% of the correlations between the Barthel Index, the Level of Rehabilitation Scale and the Functional Status Assessment Instrument were significant.

The objective of this half of the Quality of Data investigation was to examine the validity of the chosen study scales. Using concurrent, criterion-related validation techniques, these results suggested that the Barthel Index and the LORS when compared to the Functional Status Assessment Instrument (FSAI) were indeed valid for use in the Geriatric Trial.

CHAPTER VI

Summary, Implications and Limitations

Summary

In the years to come, the health care needs of the rapidly growing elderly population will have to be continually addressed in an effective and efficacy manner. Rowe (1985) pointed out that just as children are not merely young adults elderly people are not simply an older version of the mature person. The elderly individual requires special approaches and an understanding of the pathological, physiological, physical and psychosocial considerations of aging. Health care services designed with the older person in mind need to emphasis the restoration and maintenance of functional capabilities of its clientele.

Such has been the approach of the Geriatric Trial set up at the Royal Victoria Hospital. The parent study was conducted in an effort to examine the effects on the elderly patient of attaching an interdisciplinary geriatric team to the medical wards of an acute-care hospital.

In order to assure that the data collected from this controlled trial was of good quality, a detailed program of data assessment was simultaneously carried out for the duration of the Parent Investigation. It is this evaluation of data quality which has been the focus of this thesis. The main goals of this study were to estimate the reliability and the validity of the two standardized functional status indices

used in the Geriatric Trial. Given the distinct elements of these psychometric properties, two separate studies were conducted. Rater Reliability was the first aspect to be addressed. The scales used in the Geriatric Trial were the Barthel Index which examined the performance of Self Care, Continence, and Mobility and the Level of Rehabilitation Scale (LORS) which assessed Home, Outside, and Social Activities.

Prior to the start of the trial, a detailed training regime was set up to familarize all raters with the study population, the selected indices and the scoring systems. Upon completion of the training period, reproducibility testing sessions were scheduled for all 23 participants. Initially, rater scores were compared to the established Gold Standard. The results of these findings indicated that rater variability did indeed exist but that group and individual raters variations, on average, fell within acceptable norms for both study scales. The greatest overall variation was attributed to interactions between specific raters and specific patients.

Contrary to the literature and to our earlier perceptions, the Level of Rehabilitation Scale proved to be a better source of rater reliability than the more widely used Barthel Index. One reason for this discrepancy most certainly stemmed from the fact that individual patient variation was markedly limited in the areas of Self Care and Continence. Although perfect agreement was achieved between the Instructor and the Gold Standard, a five to ten point spread was seen between the Evaluators, Interpreters and the Gold Standard.

These differences resulted in neglible overall agreements for these groups. The lack of patient variability was not identified until after the video testing sessions of the twenty-three raters had been completed. As a result, what appeared to be poor agreement ratios for these two subscales was, in reality, a lack of needed patient variation for these two functional activities. Consequently, no clear statement could be taken on the degree of rater reliability for the first two subscales of the Barthel Index. In constrast, agreement levels for the remaining two Barthel subscales reached acceptable levels of concordance for all rater-gold standard comparsions.

Yet, this presence of rater consistency did not guarantee rater conformity, since systematic bias was present among the group of Evaluators. Although good agreement was expected between the Instructors and the Gold Standard for the Barthel Index, it was not anticipated among the Interpreters. These results, therefore, proved to be very interesting. One of the original goals of the study organizers was to establish a sufficiently adequate interview-training program for all raters. The belief was that greater rater variability would be generated from those individuals with the least exposure to the treatment of the elderly patient and the concepts which encompassed functional status. As a consequence, greater attention was directed towards the education of the non-professional people hired to assist in the data collection. The Interpreters predominately made up this group of people.

Results indicated that the training program had met the needs of the study's Interpreters. Furthermore, these findings seemed to demonstrate that an individual without prior medical knowledge could be appropriately trained to gather data for a controlled clinical trial. More importantly, this outcome pointed out that possessing a medically related background, did not insure superior understanding of the concepts of functional status or measurement scales. What was overlooked in this training program was the range of rater bias that could be present among individuals with similar professional experiences. In attempting to secure reliable data in future studies, subjective interpretations of even familar terms must be clarified for all study raters.

In comparison to the Barthel Index, the agreements ratios for the Level of Rehabilitation Scale provided solid evidence that good to excellent consistencies were present between all raters and the standard. Some systematic bias was found for both the Evaluators and Interpreters but this was marginal.

In general, the inter-rater reliability between the three groups of raters followed a similar pattern to the overall variation with the Gold Standard. Poor agreements again prevailed for the Barthel subscales of Self Care and Continence while fair to excellent ratios were seen for Mobility and Total Status. Overall, there seemed to be a slightly better concordance between the Interpreters and Instructors than between the other two rater combinations.

Measures of systematic bias were also very revealing.

There was a wide spread of interpretation for each of three groups particularly in the area of Total Status and to lesser extent for Mobility and Self Care. Examination of the raw data revealed that it was the Evaluators who consistently underscored these functional abilities. Once more, it was apparent that the clarity of the Barthel terms had become a key issue in this investigation.

Excellent levels of between-rater reliability continued to be attained for the Level of Rehabilitation Scale. Moreover, the rate of consistency achieved for this scale also led to excellent rater conformity. To capsulize these findings, it would appear that the Level of Rehabilitation Scale, although developed in a rehabilitation milieu, did meet the needs of an elderly population. The majority of the instrumental activities of daily living included in the scale were able to describe an elderly person's ability to function in the community.

The final impression of rater reliability was accomplished through the examination of the within-rater differences. The greatest rater variability was repeatedly found when using the Barthel Scale and the Evaluators continued to demonstrate the greatest inconsistencies. Despite the lack of patient variability for the areas of Self Care and Continence, both the Evaluators and the Interpreters were unable to agree in the estimation of functional status upon successive testing of these subscales. The Instructors, on the other hand, were consistent with the Standard and within themselves. The excellent levels of agreements for the

Level of Rehabilitation Scale achieved for the rater-standard and the between-rater comparisons continued to hold true for the intra-rater reliability.

In all, the Level of Rehabilitation Scale was found to be essentially a reliable tool for the geriatric population understudy. The Barthel Index, in constrast, demonstrated early and serious problems of rater misinterpretation as well as questionable ratios for two of the subscales. In an effort to correct this inconsistency, the user's manual was adapted and tested by means of additional patient interviews. The scores obtained from these sessions provided strong evidence that at last good to excellent rater reliability had been finally achieved for the Barthel Scale. To verify this on-going reliability of rater scores and to monitor adherence to study protocol, on-the-spot inspections were conducted over a one year period. In total, agreement ratios for both the Barthel Index and the LORS remained consistently high. Comprehension of the subscales of Self Care and Continence was no longer a problem. In sum, it can be stated that good to excellent levels of rater reliability using the Barthel Index and the LORS had not only been achieved but also maintained for the duration of the Parent Study of Geriatric Care.

Validation of the study's scales formed the second component of this investigation. The Functional Status Assessment Instrument (FSAI) by Jette (1978) was used as the criterion of comparison. This scale was chosen because of its content which addressed both issues of physical and

instrumental activities of living for an elderly population. The scores from the Barthel Index and the LORS were arranged in an ascending order, while the Jette Scale was scored in a descending form. Comparing similar components of Self Care and Mobility from both the Barthel Index and the FSAI, moderate to strong inverse relationships were obtained. The pairing of items from the LORS and FSAI were not as impressive as seen for the areas of physical status. Nevertheless, fair to moderate correlations were present for matched items of Home and Outside Activities as well as Social Interaction. Carmines and Zellers (1979) however, found that the degree of criterion-related validity depended on the extent of the correspondence between the test and the criterion. The validity coefficients obtained for the LORS-FSAI pairing appeared to reflect this finding, for the items that made-up these two scales could only be said to represent a similar concept or activity but could not be identified as like items. Cronbach (1971) also expressed an important point that should be kept in mind when attempting to establish the criterion-related validity of a scale. He has said that all validation reports carry a warning clause, insofar as the criterion selected is truly representative of the outcome to be maximized.

Implications

With the ever increasing demand to examine the effectiveness of therapeutic interventions through

well-designed controlled clinical trials, several implications from this investigation can be pertinent for other researchers and clinical personnel engaged in health care research. When designing a credible clinical trial, considerations must also be given to the development of effectual methods for assessing the data, for only through reliable and valid data can well grounded conclusions be obtained. Initially, such a program needs to address the definitions of standard procedures as well as the teaching, the training and the certification of all personnel assigned to major data collection and coding activities. As seen in this quality assessment investigation of the Geriatric Trial, the delegation of these tasks to a separate independent individual increases the objectivity of the clinical trial and liberates the principle investigators to attend to the management of the parent study. This person assumes the responsability of assuring the quality of the data collected and has the authority to implement necessary procedures to maintain high standards.

On site visits are essential to monitor adherence to study protocol and to carry out edits and data analysis designed to detect problems in the data and recording procedures. The loss of trained evaluators from a longitudinal study is inevitable; it is therefore necessary that provisions are made for the teaching and certifying of new personnel. It is also useful to have periodic meetings with the study raters to review definitions and procedures and to discuss any problems that may have been encountered.

Finally, substandard data processing and analysis procedures can be just as deleterious to the successful conduct of a trial as carelessness in the collection and the recording of the data. In essence, this form of measurement error can only be reduced through meticulous handling of the data. As Knatterud (1981) pointed out when the results of a study tend to be controversial, it becomes increasingly important that adequate quality assessment procedures have been implemented.

Limitations

In retrospect, a review of this quality assessment investigation identified three specific limitations. First, in an effort to determine the extent of rater variability in using the two study scales, video taped interviews of six different patients were presented to the twenty-three study raters. This approach was selected rather than the originally proposed Incomplete Latin Square Design in order to reduce the effects of patient learning and to eliminate patient fatigue. However, in choosing to control for these potential obstacles through the use of video equipment, an alternative limitation was introduced. The taped interviews removed the opportunity of personal contact with the patients, a comment expressed by several of the study raters. Furthermore, when ambiguous situations arose in the taped sessions, the raters stated that they found it more difficult to come to a decision since they were unable to rephrase the question to arrive at a clearer

interpretation of functional status. This limitation had been recognized by the study organizers prior to the start of the testing sessions. Weighing the importances of each restraint, it was finally felt that the video-taped approach would introduce the least bias into the assessment procedure.

Secondly, in choosing a criterion scale to establish the validation of the two geriatric indices, the Functional Status Assessment Instrument was selected. Although this scale met many of the activities covered by the Barthel Index and the Level of Rehabilitation Scale, one important area was not assessed. This was the area of Continence. In searching the literature for a comparable criterion however, the majority of the existing scales completely neglected this function. As a result, total validation of the Barthel Scale was not completed.

The third and final limitation addressed the issue of the choice of validation techniques. The criterion-related validation approach was selected for the Geriatric Trial. However, in later discussions with colleagues, the construct approach appeared to be the more appropriate format. The reason given was that when choosing a criterion for comparison, the measure chosen should represent a gold standard or a norm. This was not the case with the Functional Status Assessment Instrument (FSAI).

In summary, the evaluation of the reliability and the validity of a set of functional status scales used in a geriatric population have been presented in this study. This

quality assessment investigation appears to have reached its goal in the establishment of quality data, although certain limitations have been recognized.

Quantitative measures are increasingly being used to establish treatment objectives, to assess responses to therapeutic interventions and to develop effective treatment programs. It is, therefore, important to adhere to the established principles of measurement theory in order to assure good quality data. As Conine (1972) stated "It is axiomatic that the results of any research can be no more significant, reliable and valid than the exactness of its tools of measurement and the care with which the data are collected and processed".

Bibliography

Abramson JH: Survey Methods in Community Medicine. 2nd. ed. Churchill Livingston, New York, 1979.

Anderson AJ: Methodolgy in evaluating the quality of medical care. An annotated selected bibliography, 1962-1968. In A Guide to Medical Care Administration, Vol II: Medical Care Appraisal, Donabedian A: Ed. N.Y., American Public Health Association Pg 177-221, 1969.

Anderson S, Auger A, Hauck WW, Oakes D, Vandaele W, Weisberg HI: Statistical Methods for Comparative Studies. Techniques for bias reduction. Wiley, New York, 1980.

Anderson TP, McClure WF, Athelstan G, Anderson E, Crewe N, Arnetts L, Ferguson MB, Baldridge M, Gullickson G, Kotte FJ: Stroke rehabilitation: evaluation of its quality by assessing patient outcomes. Arch Phys Med Rehabil 59:170-173, 1978.

Applegate W, Akins D, Vander R, Thomi K, Baker MG: A geriatric rehabilitation and assessment unit in a community hospital. J Amer Geriatr Soc 31(4):206-210, 1983.

Armitage P: Statistical Methods in Medical Research. 2nd. ed. Blackwell Pub. London, 1973.

Bartko JJ: The intraclass correlation coefficient as a measure of reliability. Psychol Rep 19:3-11, 1966.

Becklake MR, Leclerc M, Strobach H, Swift J: The N closing volume test in population studies of variation and reproducibility. Amer Rev Resp Dis 111:141-147,1975.

Berger M, Bobbit RA, Pollard WE, Martin DP, Gilson BS: The sickness impact profile: Validation of a health status measure. Medical Care 14(1):57-67, 1976.

Blalock HM: The Measurement Problem, Methodology in Social Research. McGraw-Hill, New York, 1968.

Bloom BS, Soper KA: Health and medical care for the elderly and aged population: the state of the evidence. J Amer Geriatr Soc 28(10):451-455, 1980.

Blumfield S, Morris J, Sherman FT: The geriatric team in the acute care hospital: an educational and consultation modality. J Amer Geriatr Soc 30(10):660-664, 1982.

Brody SJ, Balaban DJ, Cickar G, Vermeiren JC: A diagnostic and treatment center for the aging. The Gerontologist 16:47-51, 1976.

Bromley DB: The psychology of human aging. Penguin Books, 1966.

Brook RH, Davies-Avery A, Greenfield S, Harris LJ, Lelah T, Solomon NE, Ware Jr. JE: Assessing the quality of medical care using outcome measures: an overview of the method. Medical Care 15:9, Supplement Sept. 1977.

Brook RH: Quality of care assessment: a comparison of five methods of peer review. DHEW Publications No. HRA-74-3100. Washington, D.C., U.S. Department of Health, Education and Welfare, July, 1973.

Brooks RH, Ware JE, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, Williams KN, Johnston SA: Overview of adult health status measures fielded in Rand's Health Insurance Study. Medical Care 17:1-31, 1979.

Burdock EI, Fleiss JL, Hardesty AS: A new view of inter-observer agreement. Pers Psychol 16:373-384, 1963.

Burley LE, Currie CT, Smith RG, Williamson J: Contribution from geriatric medicine within acute medical wards. B Med J 2:90, 1979.

Campbell DT, Stanley JC: Experimental and quasi-experimental designs for research. Rand McNally College Publishing Company Chicago, 1963.

Campbell LJ: Approaches to staff and student interdisciplinary team training on a geriatric evaluation unit. Paper presented at the 34th Annual Scientific Meeting of the Gerontological Society of America , Toronto Ont. Canada, 1981.

Campion EW, Jette AM, Berkman B: An interdisciplinary geriatric consultation service: a controlled trial. J Amer Geriatr Soc 31(12):792-796, 1983.

Carey RG, Posavac EJ: Program evaluation of a physical medicine and rehabilitation unit: a new approach. Arch Phys Med Rehabil 59:330-337, 1978.

Carey RG, Posavac EJ: Manual for the level of rehabilitation scale-II (Lors-II). Lutheran General Hospital Park Ridge Illinois, 1980.

Carmines EG, Zeller RA: Reliability and Validity Assessment. A Sage University Paper Pg 5-70, 1979.

Cheah KC, Baldridge JA, Beard OU: Geriatric evaluation unit of a medical service: role of a geropsychiatrist. J Gerontology 34:41-45, 1979.

Chekryn J, Roos LL: Auditing the process of care in a new geriatric unit. J Amer Geriatr Soc 28(3):107-111, 1979.

Chen MK: Health status indexes: work in progress. Health Services Research 11, 330-528, 1976.

Clarfield AM: A long-term geriatric teaching ward in an acute hospital: a three year experience. J Amer Geriatr Soc 30(7):457-465, 1982.

Clark A: Planning for adult day-care within a continuum of long-term care services. Chapter 3, Adult Day Care: A Practical Guide. Wadsworth Health Services Division, 1982.

Cochran WG, Cox GM: Experimental Designs. 2nd. edition Wiley, New York, 1957.

Colton T: Statistics in Medicine. Little Brown and Company, Boston, 1974.

Conine TA: Dilemmas of research in occupational therapy. Am J Occup Ther 26:81-84, 1972.

Cox GM: Planning Experiments. Wiley, New York, 1958.

Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 16:297-334, 1951.

Cronbach LT: Test Validation. R.L. Throndike (ed.)Educational Measurement, Washington D.C. American Council on Education Pg 443-507, 1971.

Deniston OL, Jette AM: A functional status assessment instrument: validation in an elderly population. Health Service Res 15(1): 297-334, 1980.

Donabedian A: Evaluating the quality of medical care. Milbank Mem Fund Q (Part 2) 44:166, 1966.

Donaldson SW, Wagner CC, Gresham GE: A unified ADL evaluation form. Arch Phys Med Rehabil 54:175-179, 1973.

Duke University Center for the Study of Aging and Human Development Multidimensional Functional Assessment: The Older Americans Resources and Services (OARS) Methodology. Durham, North Carolina, Duke University, 1978.

Eggert GM, Granger CV, Morris R, Pendleton SF: Caring for patients with long-term disability. Geriatrics 32:102-114, 1977.

Exton-Smith AW: Progressive patient care in geriatrics. The Lancet 1:260-263, 1962.

Farrow SC, Rablen MR, Silver CP: Geriatric admission in East London 1962-1972. Age and Aging 5:49-55, 1976.

Federer WT: Experimental Design, Theory and Application. McMillan, New York, 1953.

Feinstein AR: Clinical epidemiology. The architecture of clinical research. WB Saunders Company, 1985.

Feinstein AR: Clinical Biostatistics. Clinical Pharmacology and Therapeutics, 27(4):567-578, 1980.

Feinstein AR: Statistics versus science in the design of experiments. Clinical Biostatistics, St. Louis, The C.V. Mosby Co., 1977.

Fessel WJ, Van Brient EE: Assessing the quality of care from the medical record. N Eng J Med 286:134, 1972.

Fillenbaum G: Reliability and Validity of the OARS Multidimensional Functional Assessment Questionnaire. Center for the Study of Aging and Human Development, Durham N C, 1976.

Fillenbaum GG, Smyer MA: The development, validity and reliability of the OARS multidimensional functional assessment questionnaire. J Gerontology 36(4):428-434, 1981.

Fleiss JL, Shrout PE: The effects of measurement errors on some multivariate procedure. AJPH 67(12):1188-1191, 1977.

Fleiss JL: Measuring agreement between two judges on the presence or absence of a trait. Biometrics 31:651-659, 1975.

Freund RJ, Littell RC: SAS for Linear Models. A guide to the ANOVA and GLM Procedures. SAS Institute Inc. Box 8000 Cary, North Carolina, 1981.

Garraway WM, Akhtar AJ, Gore SM, Prescott RJ, Smith RG: Observer variation in the clinical assessment of stroke. Age and Aging, 5:233-240, 1976.

Goodnight JH: Anova. SAS User's Guide:Basic. SAS Institute Inc. Box 8000 Cary, North Carolina, 1982.

Granger CV, Albrecht GL, Hamilton BB: Outcome of comprehensive medical rehabilitation: measurement by Pulse Profile and Barthel Index. Arch Phys Med Rehabil 60:145-154, 1979.

Granger CV, Dewis LS, Peters WC, Sherwood CC, Barrett JE: Stroke rehabilitation: Analysis of repeated barthel measures. Arch Phy Med Rehabil 50:14-17, 1979.

Granger CV, Greer DS, Liset E, Coulombe J, O'Brien E: Measurement of outcome of care for stroke patients. Stroke, 6:34-41, 1975.

Granger CV, Greer DS: Functional status measurement and medical rehabilitation outcome. Arch Phys Med Rehabil 57:103-109, 1976.

Greenberg NS, Rosin AJ: Factors influencing admission or nonadmission of the aged to the hospital. J Amer Geriatr Soc 30(10):635-641, 1982.

Greenfield S, Creten S, Worthman LG, Dorey FJ, Solomon NE, Goldberg GA: Comparison of a criteria map to a criteria list in quality-of-care assessment for patients with chest pain: the relationship of each to outcome. Med Care 19:255-272, 1981.

Gresham GE, Phillips TF, Labi MLC: ADL status in stroke: relative merits of three standard indexes. Arch Phys Med Rehabil 61:355-358, 1980.

Halstead LS: Team care in chronic illnes: a critical review of the literature of the past 25 years. Arch Phys Med Rehabil 57:507-511, 1976.

Hasselkus BR, Safrit MJ: Measurement in Occupational Therapy. Amer Occup Ther 30(7):429-436, 1978.

Hays WL: Statistics. Holt, Rinehart and Winston Inc. New York, 1963.

Henderson WG, Tourtellotte WW, Potvin AR: Training examiners to administer a quantitative neurological examination for a multi-center clinical trial. Arch Phys Med Rehabil 56:289-295, 1975.

Henriksen JB: Problems in rehabilitation after age sixty-five. J Amer Geriatr Soc 24(11):510-512, 1978.

Hodkinson HM, Hodkinson I: Death and discharge from a geriatric department. Age and Aging 9:220-228, 1980.

Hodkinson HM, Jeffreys PM: Making hospital geriatric work. British Medical Journal 4:526-539, 1972.

Issac S, Michael WB: Instrumentation and Measurement. Handbook in Research and Evaluation. Edits, San Diego, 1981.

Iverson IA, Silverberg NE, Stever RC, Schoening HA: The Revised Kenny Self-Care Evaluation. Minnesota, Sister Kenny Institute, 1973.

Jette AM, Deniston OL: Inter-observer reliability of a functional status instrument. J Chron Dis 31:573-580,1978.

Jette AM: Health status indicators: Their utility in chronic disease evaluation research. J Chron Dis 333:567-579, 1979.

Jette AM: Functional capacity evaluation: an empirical approach. Arch Phys Med Rehabil 61:85-89, 1980.

Jette AM: Functional Status Index: Users Instructions. Mass General Hospital, Institute of Health Professions, Boston, Mass, 1980.

Johnson JC: Geriatric assessment units in general hospitals. J Amer Geriatr Soc 32(4):332, 1984.

Jones EW, McNitt B, McKnight E: Patient Classification for Long-Term Care: Users Manual. Department of Health Education and Welfare, HRA 75-3107, Washington, D C, 1974.

Kane RA, Kane RL: Assessing the elderly: a practical guide to measurement. Lexington Mass. Lexington Books, D.C. Health and Co., 1981.

Kane RL, Kane RA: Care of the aged: old problems in need of new solutions. Science 200:913-919, 1978.

Kane RL, Woolley FR, Gardner HJ, Snell GF, Leight EH, Castle CH: Measuring outcomes of care in an ambulatory primary care population. J Community Health 1(4):233-240, 1976.

Katz S, Ford AB, Downs TD, Adams M, Rusby DI: Effects of Continued Care: a study of chronic illness in the home. Department of Health and Welfare, DHEW, HSM 73-3010, 1972.

Katz S: Assessing self-maintaince: activities of daily-living mobility and instrumental activities of daily-living. J Amer Geriatr Soc 31(12):721-727, 1983.

Kerner JF, Alexander J: Activities of Daily Living: reliability and validity of gross vs specific ratings. Arch Phys Med Rehabil 62:161-166, 1981.

Knatterud GI: Methods of quality control and of continuous audit procedures for controlled clinical trials. Controlled Clinical Trials 1(4):327-332, 1981.

Koch GG: A general approach to the estimation of variance components. Technometrics 9:93-118, 1967.

Koch GG: Some further remarks concerning "A general approach to the estimation of variance components". Technometrics 10:551-558, 1968.

Kramer MS, Feinstein AR: Clinical biostatistics LIV. The biostatistic of concordance. Clin Pharmacol Ther 29(1):111-123, 1981.

Kuzma JW, Namerow NS, Tourtellotte WW, Sibly WA, Kurtzke JF, Rose AS, Dixon WJ: An assessment of the reliability of three methods used in evaluating the status of Multiple Sclerosis patients. J Chron Dis 21:803-814, 1969.

Kuzma JW, Tourtellotte WW, Remington RD: Quantitative clinical neurological testing II. J Chron Dis 18:303-331, 1965.

Lamont C, Sampson S, Matthias R, Kane RL: The outcome of hospitalization for acute illness in the elderly. J Amer Geriatr Soc 31(5):282-288, 1983.

Langer E, Rodin J: Effects of choice and enhanced personal responsibility for the aged. J Person Soc Psych 34:191-198, 1976.

Last JM: A dictionary of epidemiology. IEA Oxford University Press, 1983.

Lawton MP, Brody EM: Assessment of older people: self maintaining and instrumental activities of daily living. The Gerontologist 9:179-186, 1969.

Lawton MP, Moss M, Fulcomer M, Kleban MH: A research and service oriented multilevel assessment instrument. J Gerontology 37(1):91-99, 1982.

Lefton E, Bonstelle S, Dermot Frengley J: Success with an inpatient geriatric unit: a controlled study of outcome and follow-up. J Amer Geriatr Soc 31:3, 149-155, 1983.

Lefton E, Lefton M: Health care and treatment for the chronically ill: toward a conceptual framework. J Chron Dis 32:339-344, 1979.

Lichtenstein H, Winograd CH: Geriatric consultation: a functional approach. J Amer Geriatr Soc 32(5):356-361, 1984.

Linn MW, Linn BS: The rapid disability rating scale II. J Am Geriatr Soc 30(6):378-382, 1982.

Loewenson RB, Bearman JE, Resch JA: Reliability of measurement for studies of Cerebrovascular Atheroslerosis. Biometrics 28:557-569, 1972.

Mahoney FI, Barthel DW: Functional evaluation: the Barthel Index. M D State Med J 14:61-65, 1965.

McAuliffe WE: Measuring the quality of medical care: process versus outcome. Milbank Mem Fund Q Health and Society 57(1)118-152, 1979.

Moser CA, Kalton G: Scaling Methods. Survey Methods in Social Investigation. London: Heinemann Educational Books, 350-377, 1971.

Nobrega FT, Morrow GW, Smoldt RK, Offord KP: Quality assessment in hypertension: Analysis of proces and outcome methods. N Eng J Med 296:145, 1977.

Nunnally JC: Educational Measurement and Evaluation. McGraw-Hill, New York, 1964.

Nunnally JC: Validity in Psychometric Theory. McGraw-Hill, New York, 1978.

O'Brien CL: Adult Day Care: A Practical Guide. Wadsworth Health Services Division, 1982.

O'Brien TD, Joshi DM, Warren EW: No apology for geriatrics. British Medical Journal 4:277-280, 1973.

Parry F: Physical rehabilitation of the old, old patient. J Amer Geriatr Soc 31(8):482-484, 1983.

Patrick DL, Bush JW, Chen MK: Toward an operational definition of health. J Health and Soc Behav. 14:6-23, 1973.

Patterson C, Crescenzi C: Hospital use by the extremely elderly (nonagenarians). J Amer Geriatr Soc 32(5):350-352, 1984.

Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S: Measurement of functional activities in older adults in the community. J Geron 37(3):323-329, 1982.

Pfeiffer E: Mult-dimensional Functional Assessment: the OARS Methodology. Center for the Study of Aging and Human Development, Durham, N C, 1976.

Pfeiffer E: A short portable mental status questionnaire for the assessment of Organic Brain Deficit in elderly patients. J Amer Ger Soc 20(3):10, 433-441, 1975.

Potvin AR, Tourtellotte WW, Henderson WG, Synder MS: Quantitative examination of neurological function: reliability and learning effects. Arch Phys Med Rehabil 56:438-442, 1975.

Potvin AR, Tourtellotte WW: The neurological examination: advancements in its quantification. Arch Phys Med Rehabil 56:425-437, 1975.

Reynolds WJ, Rushing WA, Miles DL: The validation of a function status index. J Health and Soc Behav. 15:271-288, 1974.

Robertson D, Rockwood K: Outcome of hospital admission of the very elderly. J Amer Geriatr Soc 30:101-104, 1982.

Romm FJ, Hulka BS, Mayo F: Correlated of outcomes in patients with congestive heart failure. Med Care 14:765, 1976.

Rombout MK: Hospitals and the Elderly: Present and Future Trends. Health and Welfare Canada, 1975.

Rothberg JS: The rehabilitation team: future direction. Arch Phys Med Rehabil 62(8):407-409, 1981.

Rowe JW: Health care of the elderly. N Engl Med J 32(13):827-835, 1985.

Rubenstein L: The clinical effectiveness of multidimensional geriatric assessment. J Amer Geriatr Soc 31:758-762, 1983.

Rubenstein L, Abrass IB, Kane RL: Improved care for patient on a new geriatric evaluation unit. J Amer Geriatr Soc 29(11):531-536, 1981.

Rubenstein L, Kane RL: Geriatric assessment units in general hospitals. J Amer Geriatr Soc 32(4):331, 1984.

Sall JP, Delong DM: Correlation Procedures. SAS User's Guide:Basic. SAS Institute Inc. Box 8000 Cary, North Carolina, 1982.

Sarno JE, Sarno MT, Levita E: Functional Life Scale. Arch Phys Med Rehabil 54:214-220, 1973.

Schuman JE, Beattie EJ, Steed DA, Gibson JE, Merry GM, Campbell WD, Kraus AS: The impact of a new geriatric program in a hospital for the chronically ill. Canadian Medical Association Journal 118:639-645, 1978.

Schuman JE, Beattie EJ, Steed DA, Merry GM, Kraus AS: Geriatric patients with and without intellectual dsyfunction: effectiveness of a standard rehabilitation program. Arch Phys Med Rehabil 62:612-618, 1981.

Sherwood S, Morris JN: A study of the effects of emergency alarm and response system for the aged: a final report. Boston, Massuchuetts, Hebrew Rehabilitation Center for the Aged,1980.

Sherwood SJ, Morris J, Mor V, Gutkin C: Compendium of measures for describing and assessing long term care populations. Boston: Hebrew Rehabilitation Center for Aged, 1977.

Silver CP, Uberi SJ: Prognosis of patients admitted to a geriatric unit. Geront Clin 7:348-357, 1965.

Sloane PD: Nursing home candidates: hospital inpatient trial to identify those appropriately assignable to less intensive care. J Amer Geriatr Soc 28(10):511-514, 1980.

Starfield B, Scheff D: Effectiveness of pediatric care: The relationship between process and outcome. Pediatrics 49:547, 1972.

Statistics Canada: Population Projections for Canada, Provinces and Territories, Catalogue 91-520, 1984-2006.

Stewart A, Ware JE, Brook RH: The meaning of health: understanding functional limitations. Medical Care 15:939-952, 1977.

Stone LO, Fletcher S: Aspects of Population Aging in Canada. A chartbook. National Advisory Council on aging. Statistics Canada, 1981.

Teasdale TA, Luchi RJ, Shuman L, Snow E: Geriatric assessment units in general hospitals. J Amer Geriatr Soc 32(4):333, 1984.

Teasdale TA, Shuman L, Snow E, Luchi RI: A comparison of placement outcomes of geriatric cohorts receiving care in a geriatric assessment unit and on general medicine floors. J Amer Geriatr Soc 31(9):529-534, 1983.

Tourtellotte WW, Haerer AF, Simpson JF, Kuzma JW, Sikorski J: Quantitative clinical neurological testing I. A study of a battery of tests designed to evaluate in part the neurological function of patients with Multiple Sclerosis and its use in a clinical trial. Annuals New York Academy of Science, 122:480-505, 1965.

Vogel RJ, Palmer HC: Long-Term Perspectives from Research and Demonstrations. Health Care Financing Administration, U.S. Dept. of Health and Human Services, 1983.

Williams RGA, Johnston M, Willis LA, Bennett AE: Disability: A model and measurement technique. British Journal of Preventive Social Medicine 30(2):71-78, 1976.

Wolinsky FD, Coe RM, Miller DK, Prendergast JM: Measurement of global and functional dimensions of health status in the elderly. J Gerontol 39(1):88-92, 1984.

Wolinsky FD, Uusman ME: Toward comprehensive health status measures. The Sociological Quarterly 21:607-621, 1980.

Wood-Dauphinee S, Clarfield AM: Evaluating Geriatric Team Care Programs: Why and How Clinical Gerontologist 3(3):23-34, 1985.

Wylie CM: Gauging the response of stroke patients to rehabilitation. J Amer Geriatr Soc 15:797-805, 1967.

Yates F: Incomplete Latin Squares. J Agr Sci 26:301-315, 1936.

Youden WJ: Use on incomplete block replications in estimating tobacco-mosaic virus. Contr Boyce Thompson Institute, 1937.

Appendices

Appendix I A

Informed Consent

McGill University

Royal Victoria Hospital

I _____ agree to participate in a video taped interview to be used in a Geriatric Study being conducted by the Department of Medicine of the Royal Victoria Hospital.

I have been told that the objective of this taped interview is to teach the individuals who are evaluating patients in this study, the procedures for conducting a proper assessment.

I understand that the taped interview will be 30-40 minutes in length and that I will be asked 3 sets of questions pertaining to my mobility, my activities of daily living, and my general independence.

I am aware that my anonymity can not be preserved because of the nature of the taped interview.

I have been told that this project has been approved by the research committee of the Constance Lethbridge Center.

I authorize McGill University and the Royal Victoria Hospital to use this video tape for scientific and educational purposes.

Name
Date
Signature
Witness

Appendix IB

Formule de Consentement

Université McGill

Hôpital Royal Victoria

Je soussigne _____, consens à participer à une entrevue dont l'enregistrement video sera utilisé pour une étude conduite par le service de Médicine en Geriatrie de l' Hôpital Royal Victoria.

Il est entendu que l'objectif de cette enregistrement est d'enseigner aux personnes qui evaluent les patients comment conduire une evaluation pour les personnes dans cette étude.

Il est entendu que la durée de l'entrevue enreistrée sera de 30-40 minutes et qu'on me demandera des questions concernant ma mobilité, mes activités quotidiennes et mon indépendance en général.

Je comprends que mon anomymat ne peut pas être preservé en raison de l'entrevue enregistrée.

On m'a inform(e) que ce projet a été approuvé par le comité de recherche du Centre Réadaptation Constance Lethbridge.

J'autorise(e) l'Université McGill et l' Hôpital Royal Victoria d'utiliser cet enregristement sonore pour des fins scientifiques et éducatives.

Nom
Date
Signature
Témoin

BARTHEL SCALE

PATIENT NAME: _____ WARD: _____

Examiner _____ Date: _____

Examination: Admission ☐
2 weeks post admission ☐
1 month post admission ☐
3 months post admission ☐
6 months post admission ☐

Study No: _____

SELF CARE

		Self	Some Aid	Can't Do	Score	Sub-section Total
1.	Drinks from cup	4	0	0		
2.	Eating	6	3	0		
3.	Dress-upper body	5	3	0		
4.	Dress-lower body	7	4	0		
5.	Put on brace	0	-2	0		
6.	Grooming	5	0	0		
7.	Washing	6	0	0		
8.	Bladder control	10	5	0		
9.	Bowel control	10	5	0		

MOBILITY

10.	Transfer chair	15	7	0		
11.	Transfer toilet	6	3	0		
12.	Tub or shower	1	0	0		
13.	Walks 50 yds.	15	10	0		
14.	Stairs	10	5	0		
15.	Wheeling if not walking	5	0	0		
		100	TOTAL			

LORS RATING SHEET

Name of Patient _____ Informant _____

Evaluator _____ Date _____

Evaluation Pre-Admission ☐ Place _____
2 Weeks Post-Admission ☐
1 Month Post-Admission ☐ Study Number _____
3 Months Post-Admission ☐
6 Months Post-Admission ☐

Home Activities

30. Prepares simple foods and drinks (e.g. juice, toast, coffee) ***
0< Does not prepare simple foods
2< Deficient performance, prepares such foods infrequently, needs much encouragement, and/or has frequent accidents
4< Normal efficiency in preparing simple foods

31. Performs light housekeeping chores (e.g. meals, dishes, dusting) ***

32. Performs heavy housekeeping chores (e.g. floor and window washing) **

33. Performs odd jobs in or around the house (e.g. gardening, minor ** repairs, mending, sewing)

These activities are rated following the pattern developed for activity 30.

34. Engages in individual pastimes (e.g. selective TV viewing, rea- *** ding, knitting, collecting, etc.)
0< Does not engage in individual pastimes
2< Patient experiences difficulties with pastimes or hobbies because of frequent accidents, fatigue, depression
4< Normal degree of individual pastimes performed

35. Manipulates telephone and/or television (e.g. dials number, chan- *** ges stations)
0< Does not manipulate these items
2< Manipulates these items only some of the time because of accidents, or because patient fatigues easily
4< Normal use of these items

Outside Activities

36. Engages in simple outside activities (e.g. walks, car rides, sit- *** ting on porch)
0< Does not engage in these activities
2< Will engage in simple out door activities only with encouragement
4< Attempts simple outdoor activities independently; accepts or seeks help in doing the activity.

37. Does shopping and other errands (e.g. food, clothes, banking) ***
0< Does not go shopping or do other errands
2< Will attempt these activities only with encouragement
4< Attempts these sorts of activities on his own without the encouragement of others or seeks the assistance of others

Attends spectator events (e.g. theater, concerts, movies, sports)
0< Does not attend spectator sports
2< Will attend spectator events only when encouraged to do so
4< Attempts to attend such events independently or seeks help to make attendance possible.

39. Uses transportation accompanied (e.g. auto, cab, train, plane) ** 0< Does not attempt to go anywhere transportation would be required 2< Will attempt to use transportation if encouraged 4< Accepts or seeks help in using transportation	
40. Uses transportation independently (Rate "Not Applicable" if 39 *** was 0.) 0< Does not use transportation independently 2< Will use transportation independently if encouraged 4< Uses transportation independently, in a normal fashion	
41. Takes longer trips (5 hours) accompanied * 0< Does not take longer trips 2< Will attempt longer trips with encouragement 4< Attempts longer trips with encouragement, accepts or seeks help in going on longer trips	
42. Takes longer trips (5 hours) independently (Rate "Not Applicable" * if 41 was 0.) 0< Does not take longer trips independently 2< Will attempt a longer trip with encouragement 4< Attempts longer trips unaccompanied and without encouragement	
<u>Social Interaction</u> 43. Participates in games with other people (e.g. cards, chess, chec- ** kers). Do not rate quality of patient's skill. 0< Does not participate in games but did before the illness. 2< Will participate in games if encouraged 4< Initiates games with other persons	
44. Participates in home social activities (e.g. family gatherings, *** visits to friends, parties) 0< Does not participate in these activities 2< Participates in these activities only with encouragement or for only a very brief time or only in a very limited fashion 4< Patient participates in these activities without encouragement, invites or asks spouse to invite relatives and friends in	
45. Attends social functions outside of home (e.g. home of friend, ** dining at a restaurant) 0< Does not participate in these activities 2< Participates in these activities only with encouragement or for only a very brief time or only in a very limited fashion 4< Attemps to participate in social functions outside of the home on patient's own or asks to be helped to perform such activities	
46. Goes to church or synagogue * 0< Does not go to church or synagogue 2< Attends only if encouraged by others 4< Attends without encouragement and/or asks assistance to go	

TOTAL

KEY: ASSISTANCE: 1 = independent; 2 = uses devices; 3 = uses human assistance; 4 = uses devices & human assistance; 5 = unable to do

PAIN: 0 ⟶ 7, where 0 = no pain and 7 = extremely severe pain

DIFFICULTY: 0 ⟶ 7, where 0 = not difficult and 7 = extremely difficult

Time Frame - on the average during the past (7) days

ACTIVITY	ASSISTANCE (1⟶5)	PAIN (0⟶7)	DIFFICULTY (0⟶7)	COMMENTS
Mobility				
- walking inside				
- climbing up stairs				
- transferring to & from toilet				
- getting in & out of bed				
- driving a car				
Personal Care				
- combing hair				
- putting on pants				
- buttoning clothes				
- washing all parts of the body				
- putting on shoes/slippers				
Home Chores				
- vacuuming a rug				
- reaching into high cupboards				
- doing laundry				
- washing windows				
- doing yardwork				
Hand Activities				
- writing				
- opening containers				
- turning faucets				
- cutting food				
Vocational				
- performing all job responsibilities				
Avocational				
- performing hobbies requiring hand work				
- attending church				
- socializing with friends and relatives				

APPENDIX 5

DIFFERENCES BETWEEN THE GOLD STANDARD AND GROUP I (EVALUATORS) IN USING THE BARTHEL INDEX IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	95	84	80	99	89	94

Video Sessions

	I II	I II	I II	I II	I II	I II
Group I (Evaluators)						
1	95-95	84-84	80-80	85-90	70-70	89-80
2	95-95	90-74	80-80	88-94	79-79	73-88
3	95-95	89-89	80-80	99-99	89-89	94-94
5	100-95	89-89	80-70	90-94	75-89	93-86
7	95-95	75-74	80-80	99-94	63-89	89-88
8	95-95	89-84	80-80	99-93	75-89	89-83
9	90-95	89-90	80-80	99-99	91-94	99-99
10	90-90	89-74	80-79	99-94	79-67	84-83
11	90-90	74-90	80-80	71-94	72-67	89-84
12	100-100	88-88	80-80	99-99	64-92	99-97
13	80-90	89-74	80-80	93-84	88-77	99-99
14	95-95	84-84	80-80	76-100	79-79	85-94
15	90-95	84-85	80-80	89-94	79-78	85-94
16	90-90	74-74	80-80	89-99	77-77	89-79
Groups Means						
VideoI	92.8	84.7	80.0	91.0	77.1	89.7
VideoII	93.2	82.3	79.2	94.7	81.1	89.1
Group Bias *						
VideoI	-2.1	0.78	0.00	-7.9	-11.8	-4.2
VideoII	-1.7	-1.6	-0.78	-4.2	- 7.8	-4.8
Percentage of Difference						
VideoI	2.2	0.01	0.00	7.9	13.3	4.4
VideoII	1.8	1.90	0.01	4.2	8.8	5.1

* Group Bias is the measure of difference between the Evaluator Scores and the true scores (Gold Standard)

APPENDIX 6

DIFFERENCES BETWEEN THE GOLD STANDARD AND GROUP II (INTERPRETERS) IN USING THE BARTHEL INDEX IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	95	84	80	99	89	94

	Video Sessions											
	I	II	I	II	I	II	I	II	I	II	I	II
Group II (Interpreters)												
17	95	95	85	85	70	80	99	99	75	70	94	78
18	94	95	89	89	80	80	94	99	69	86	95	89
19	100	100	90	90	80	80	99	99	88	89	94	99
20	100	90	89	90	80	80	95	94	86	88	94	89
21	95	95	83	95	80	80	99	94	76	79	94	89
Group Means												
VideoI	96.8		87.5		78.0		97.2		78.8		94.2	
VideoII	95.0		89.8		80.0		97.00		82.4		88.8	
Group Bias *												
VideoI	1.8		3.2		-2.0		-1.8		-10.2		2.0	
VideoII	0.0		5.8		0.0		-2.0		- 6.6		-5.2	
Percentage of Differences												
VideoI	1.8		3.8		2.5		1.8		11.4		2.1	
VideoII	0.0		6.9		0.0		2.0		7.4		5.5	

* Group Bias is the measure of difference between the Interpreters scores and the true scores (Gold Standard)

APPENDIX 7

DIFFERENCES BETWEEN THE GOLD STANDARD AND GROUP III (INSTRUCTORS) IN USING THE BARTHEL INDEX IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	95	84	80	99	89	94

	Video Sessions											
	I	II	I	II	I	II	I	II	I	II	I	II
Group III (Instructors)												
22	95	100	84	89	80	80	100	99	79	94	94	99
23	95	100	89	89	80	80	99	99	79	94	99	99
24	95	100	89	89	80	80	99	99	80	79	94	99
25	95	95	89	84	80	80	99	99	89	89	94	94
Group Means												
VideoI	95.0		87.7		80.0		99.2		65.4		95.2	
VideoII	98.7		87.7		80.0		99.0		89.0		97.7	
Group Bias *												
VideoI	0.0		3.7		0		2.5		-23.6		1.2	
VideoII	3.7		3.7		0		0.0		0.0		3.7	
Percentage of Differences												
VideoI	0.0		4.4		0.0		0.002		27.0		1.3	
VideoII	3.9		4.4		0.0		0.000		0.0		3.9	

* Group Bias is the measure of difference between the Instructors scores and the true scores (Gold Standard)

APPENDIX 8

DIFFERENCES BETWEEN THE GOLD STANDARD AND GROUP I (EVALUATORS) IN USING THE LORS IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	78	28	67	57	63	40
	Video Sessions					
	I II	I II	I II	I II	I II	I II
Group I (Evaluators)						
1	68-71	25-25	50-47	38-34	41-41	31-31
2	76-76	31-31	67-63	44-41	50-53	38-38
3	76-78	38-30	70-70	56-53	59-59	44-38
5	73-71	29-28	54-43	58-43	50-63	36-40
7	76-73	28-23	77-70	47-50	59-56	53-41
8	84-78	68-27	63-59	47-47	56-59	50-44
9	88-78	34-27	63-67	53-47	59-50	47-50
10	59-56	28-23	44-41	38-35	50-41	38-41
11	74-78	30-31	69-56	35-41	62-59	47-41
12	76-76	43-27	66-60	41-47	47-62	38-44
13	71-68	26-33	72-66	38-32	62-50	44-47
14	71-64	26-34	54-57	47-41	56-44	50-47
15	78-76	30-27	66-56	32-35	47-47	38-44
16	71-78	28-28	67-73	47-44	59-56	44-44
Group Means						
VideoI	74.3	33.1	63.0	44.3	54.0	42.7
VideoII	72.9	28.0	59.0	42.1	52.8	42.1
Group Bias *						
VideoI	-3.6	5.10	-4.0	-12.6	- 8.9	2.7
VideoII	-5.1	0.14	-7.8	-14.8	-10.1	2.1
Percentage of Difference						
VideoI	4.6	18.2	5.9	22.0	14.0	6.7
VideoII	6.5	0.01	11.6	26.0	16.0	5.2

* Group Bias is the measure of difference between the Evaluators scores and the true scores (Gold Standard)

APPENDIX 9

DIFFERENCE BETWEEN THE GOLD STANDARD AND GROUP II (INTERPRETERS) USING THE LORS IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	78	28	67	57	63	40

	Video Sessions											
	I	II	I	II	I	II	I	II	I	II	I	II
Group II (Interperters)												
17	63	65	33	31	67	60	42	44	56	50	47	41
18	68	74	29	43	69	69	44	41	62	65	44	34
19	82	76	33	20	70	67	50	44	56	56	44	44
20	71	59	23	28	59	62	44	44	41	50	38	41
21	78	78	35	27	63	72	38	41	50	47	44	47
Group Means												
VideoI	72.4		38.2		65.6		43.6		53.0		43.4	
VideoII	70.4		29.8		66.0		42.8		53.6		41.4	
Group Bias *												
VideoI	-5.6		10.25		-1.4		-13.6		-10.0		3.5	
VideoII	-7.6		1.80		-1.0		-14.2		- 9.4		1.4	
Percentage of Difference												
VideoI	7.0		36.6		2.0		23.8		15.8		8.7	
VideoII	9.7		6.4		1.4		24.9		14.9		3.5	

* Group Bias is the measure of difference between the Interpreters scores and the true scores (Gold Standard)

APPENDIX 10

DIFFERENCES BETWEEN THE GOLD STANDARD AND GROUP III (INSTRUCTORS) IN USING THE LORS IN TWO VIDEO SESSIONS

	Patients					
	1	2	3	4	5	6
Gold Standard	78	28	67	57	63	40

	Video Sessions											
	I	II	I	II	I	II	I	II	I	II	I	II
Group III (Instructors)												
22	71	78	23	20	70	70	35	38	44	56	38	44
23	78	78	37	28	64	68	32	41	56	53	41	53
24	81	74	46	34	70	60	44	47	50	56	44	41
25	81	82	27	27	71	72	46	47	47	53	47	47
Group Means												
VideoI	77.7		33.2		68.7		39.2		49.2		42.5	
VideoII	78.0		27.2		67.5		43.2		54.5		46.2	
Group Bias *												
VideoI	-0.25		5.25		1.75		-17.75		-13.75		2.50	
VideoII	0.00		-0.75		0.50		-13.75		- 8.50		6.25	
Percentage of Difference												
VideoI	0.003		18.7		2.60		31.0		21.8		6.2	
VideoII	0.000		2.6		0.01		24.0		13.4		15.6	

* Group Bias is the measure of difference between the Instructors scores and the true scores (Gold Standard)

APPENDIX 11

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and Gold Standard for the Self Care Subscale of the Barthel Index

Barthel Index

Self Care

Sources of Variation	Expected Mean Squares Groups I	II	III	Coefficient of Variation Groups I	II	III
+ Var(groups)	0.50	0.61	0	2.20	0.7	0
++ Var(rater within group)	0.14	0.21	0	1.00	1.4	0
* Var(patients)	0.86	0.10	0	3.00	1.0	0
** Var(rater x patient within group)	1.90	2.70	0	4.20	5.0	0
*** Var(video+ random error)	0.04	0.08	0	0.58	0.8	0

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation of patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 12

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Continence Subscale of the Barthel Index

Barthel Index

Continence

Sources of Variation	Expected Mean Squares Groups I	II	III	Coefficient of Variation Groups I	II	III
+ Var(groups)	0.00	0.00	0	0.00	0.0	0
++ Var(rater within group)	0.09	0.31	0	1.50	2.9	0
* Var(patient)	0.00	0.37	0	0.00	3.1	0
** Var(rater x patient within group)	0.79	2.20	0	4.40	7.5	0
*** Var(video+ random error)	0.03	0.17	0	0.83	2.1	0

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 13

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Mobility Subscale of the Barthel Index

Barthel Index

Mobility

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I	II	III	I	II	III
+ Var(groups)	1.10	0.0	0.0	9.5	0.0	0.0
++ Var(rater within group)	7.10	2.5	1.2	7.6	4.0	2.8
* Var(patients)	37.60	42.9	54.3	17.7	17.7	19.5
** Var(rater x patient within group)	10.40	4.0	3.0	2.3	1.7	4.6
*** Var(video+ random error)	0.64	0.44	3.1	2.3	1.7	4.6

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 14

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Home Activities Subscale of the LORS

LORS

Home Activities

Sources of Variation	Expected Mean Squares Groups I	II	III	Coefficient of Variation Groups I	II	III
+ Var(groups)	0.72	0.69	0.00	6.1	5.9	0.0
++ Var(rater within group)	1.10	0.79	1.75	7.7	6.4	9.2
* Var(patients)	22.80	21.90	22.30	34.8	33.6	32.9
** Var(rater x patient within group)	3.00	1.70	3.40	12.6	9.4	12.9
*** Var(video+ random error)	0.07	0.11	0.36	1.9	2.3	4.1

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 15

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Outside Activities Subscale of the LORS

LORS

Outside Activities

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I	II	III	I	II	III
+ Var(groups)	0.0	0.0	0.0	0.0	0.0	0.0
++ Var(rater within group)	1.8	0.0	0.0	11.4	0.0	0.0
* Var(patients)	29.5	40.4	44.4	46.0	52.7	56.0
** Var(rater x patient within group)	7.4	4.8	2.2	23.0	18.3	13.4
*** Var(video+ random error)	0.07	0.30	1.0	2.2	4.5	8.4

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 16

Expected Mean Squares and Coefficients of Variation for the Three Groups of Raters and the Gold Standard for the Social Activities Subscale of the LORS

LORS

Social Activities

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I	II	III	I	II	III
+ Var(groups)	0.0	0.0	0.02	0.0	0.0	0.0
++ Var(rater within group)	0.0	0.0	0.00	0.0	0.0	0.0
* Var(patients)	19.8	20.9	25.60	58.9	60.3	65.4
** Var(rater x patient within group)	1.5	1.1	1.10	16.6	13.8	13.6
*** Var(video+ random error)	0.16	0.07	0.01	5.2	3.6	1.7

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 17

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Self Care Subscale of the Barthel Index

Barthel Index

Self Care

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I + II	I + III	II + III	I + II	I + III	II + III
+ var(groups)	0.0	0.78	0.56	0.0	2.7	2.30
++ var(rater within group)	0.17	0.09	0.16	1.2	0.9	1.22
* var(patients)	0.78	0.54	0.03	2.7	2.2	0.49
** var(rater x patient within group)	2.18	1.70	1.33	4.6	4.0	4.10
*** var(video+ random error)	0.06	0.02	0.03	0.73	0.49	0.56

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 18

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the the Continence Subscale of the Barthel Index

Barthel Index

Continence

Sources of Variation	Expected Mean Squares Groups I + II	I + III	II + III	Coefficient of Variation Groups I + II	I + III	II + III
+ var(groups)	0.04	0.0	0.13	1.0	0.0	1.8
++ var(rater within group)	0.39	0.08	0.22	1.9	0.0	1.8
* var(patients)	0.11	0.0	0.15	1.6	0.0	1.9
** var(rater x patient within group)	1.32	0.65	1.49	5.8	4.0	6.2
*** var(video+ random error)	0.0	0.02	0.08	0.0	0.69	1.4

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 19

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Mobility Subscale of the Barthel Index

Barthel Index

Mobility

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I + II	I + III	II + III	I + II	I + III	II III
+ var(groups)	5.07	9.62	0.16	6.3	8.7	1.10
++ var(rater within group)	5.32	5.42	0.31	6.5	6.6	1.48
* var(patients)	38.70	41.60	49.58	17.7	18.2	18.80
** var(rater x patient within group)	9.70	9.40	4.48	8.8	8.6	5.60
*** var(video+ random error)	0.18	1.33	0.29	1.1	3.2	1.42

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 20

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Home Activities Subscale of the LORS

LORS

Home Activities

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I + II	I + III	II + III	I + II	I + III	II + III
+ var(groups)	0.00	0.15	0.0	0.0	0.0	2.8
++ var(rater within group)	0.91	0.95	0.47	7.0	7.0	4.8
* var(patients)	22.44	22.93	22.14	34.7	34.8	33.8
** var(rater x patient within group)	2.99	3.17	2.72	12.6	12.9	11.8
*** var(video+ random error)	0.10	0.15	0.31	2.3	2.8	3.9

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 21

Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Outside Activities Subscale of the LORS

LORS

Outside Activities

Sources of Variation	Expected Mean Squares			Coefficient of Variation		
	Groups			Groups		
	I + II	I + III	II + III	I + II	I + III	II + III
+ var(groups)	0.00	0.00	0.00	0.00	0.00	0.0
++ var(rater within group)	1.79	1.85	0.06	11.32	11.56	2.0
* var(patients)	31.07	31.43	40.61	47.10	47.60	53.5
** var(rater x patient within group)	7.13	6.81	3.99	22.50	22.10	16.7
*** var(video+ random error)	0.001	0.003	0.86	0.29	0.46	6.1

\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error

APPENDIX 22

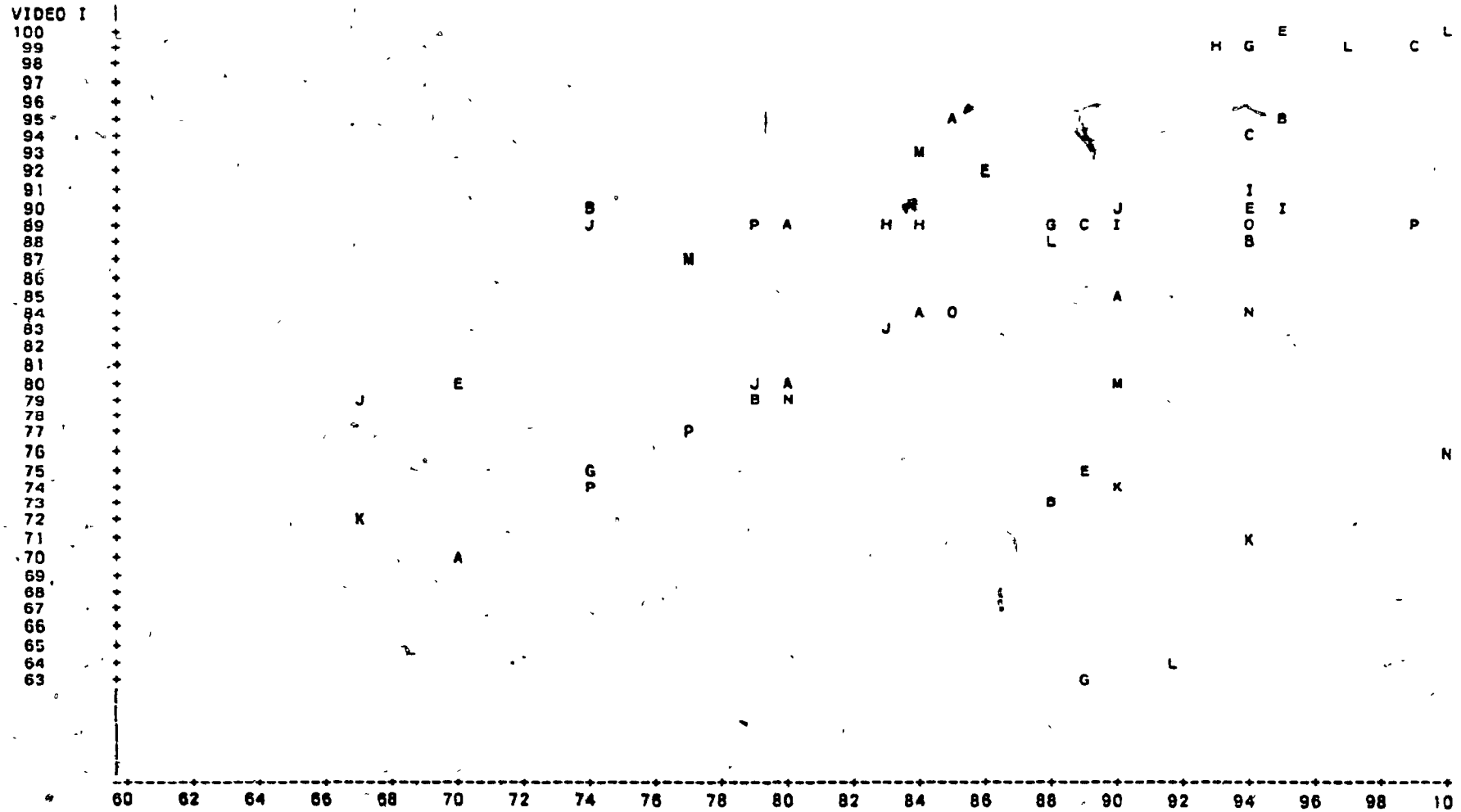
Expected Mean Squares and Coefficients of Variation When Comparing the Three Groups of Raters for the Social Activities Subscale of the LORS

LORS

Social Activities

Sources of Variation	Expected Mean Squares: Groups I + II	Groups I + III	Groups II + III	Coefficient of Variation: Groups I + II	Groups I + III	Groups II + III
+ var(groups)	0.00	0.00	0.00	0.0	0.0	0.0
++ var(rater within group)	0.18	0.17	0.02	5.6	5.4	1.9
* var(patients)	18.98	20.05	21.14	57.5	58.7	59.7
** var(rater x patient within group)	1.25	1.35	0.71	14.7	15.2	10.9
*** var(video+ random error)	0.16	0.09	0.01	5.3	3.8	1.4

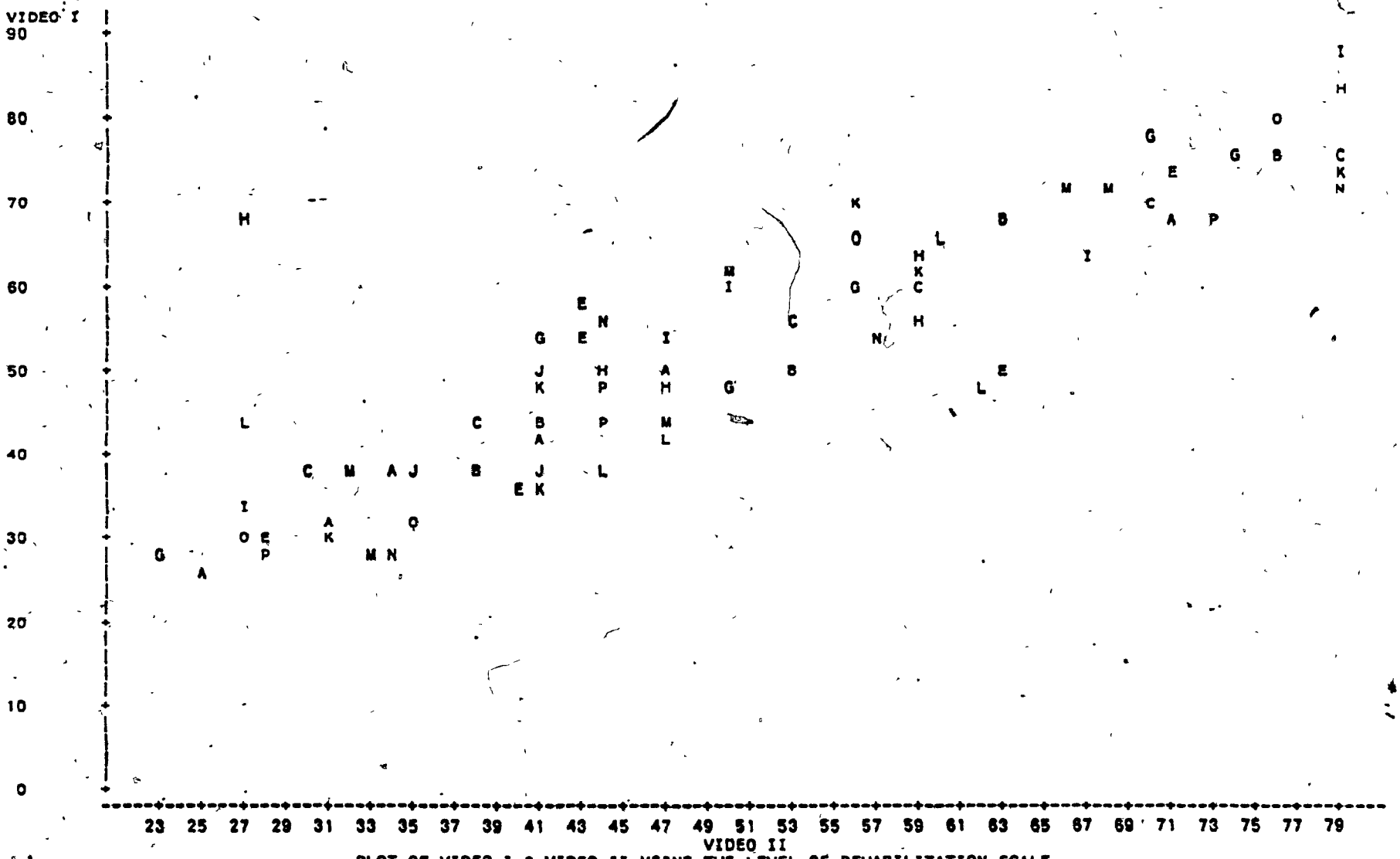
\+ var(groups)=variation between groups
++ var(r in group)=variation of rater in group
* var(pts)=variation between patients
** var(r-pts in group)=variation of rater by patient in group
*** var(video+error)=variation of video session + random error



VIDEO II

PLOT OF VIDEO I * VIDEO II USING THE BARTHEL INDEX
GROUP I=14 EVALUATORS (A TO P)

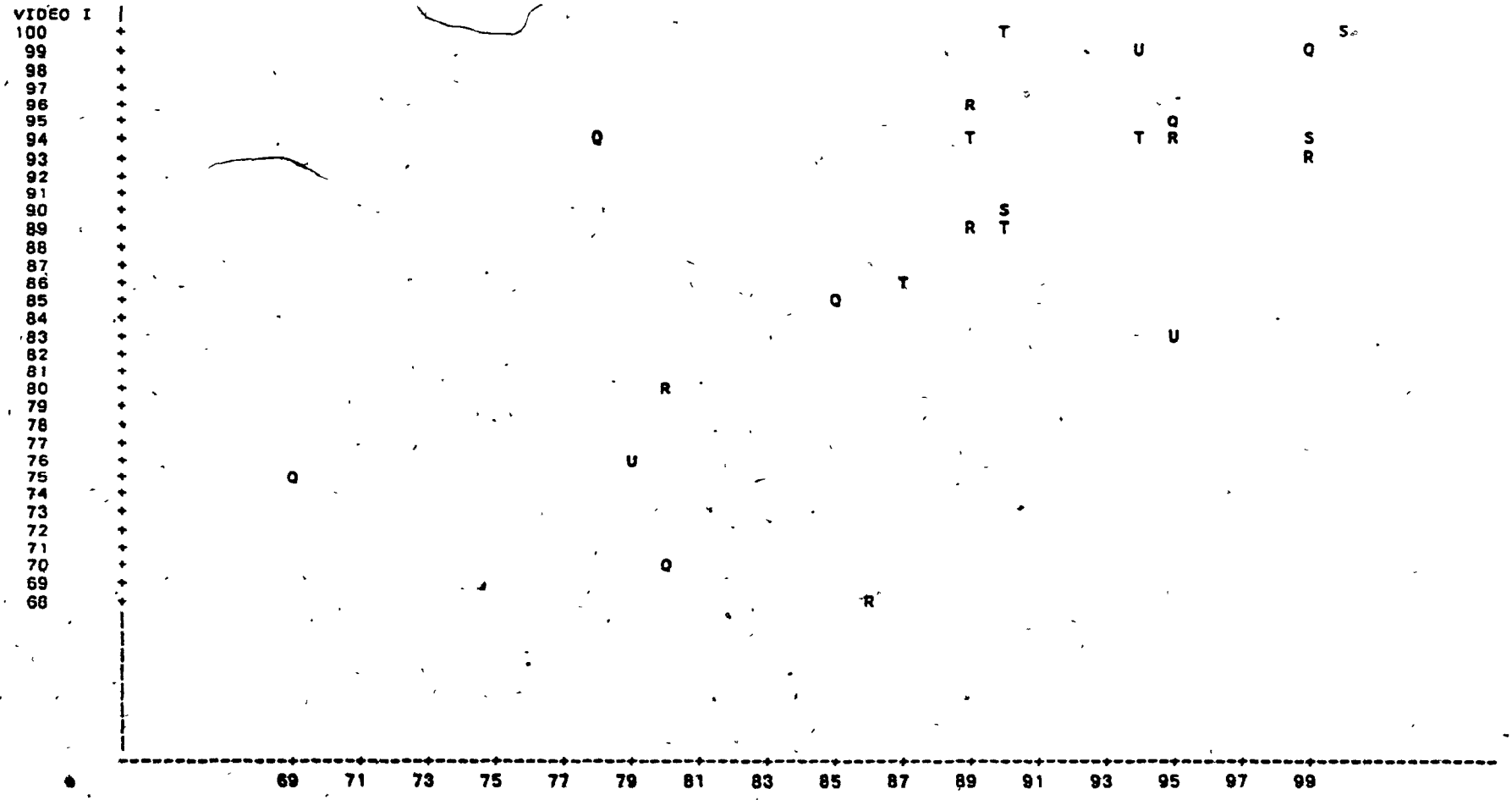
APPENDIX 23



PLOT OF VIDEO I * VIDEO II USING THE LEVEL OF REHABILITATION SCALE

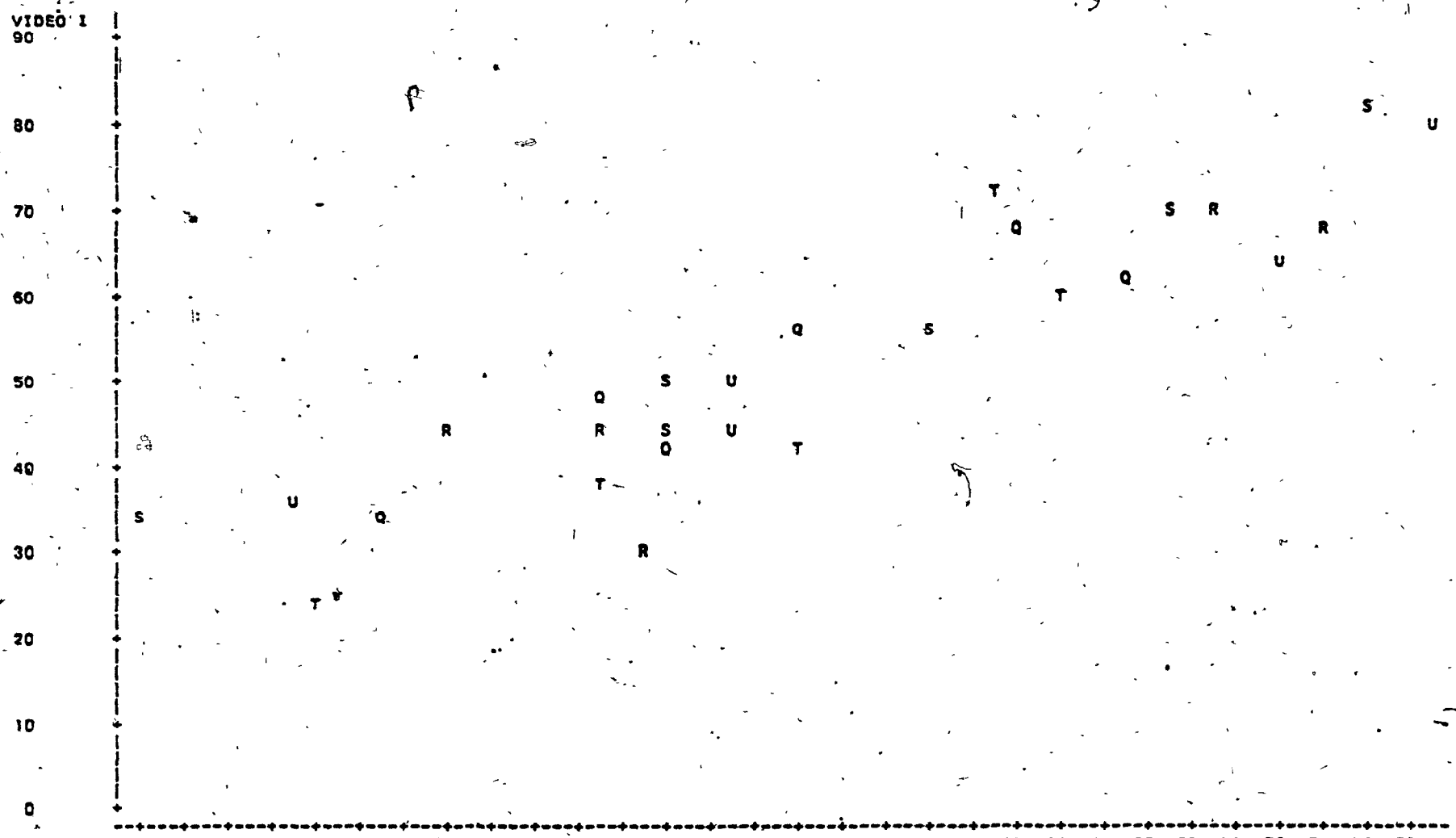
APPENDIX 24

GROUP I=14 EVALUATORS (A TO P)



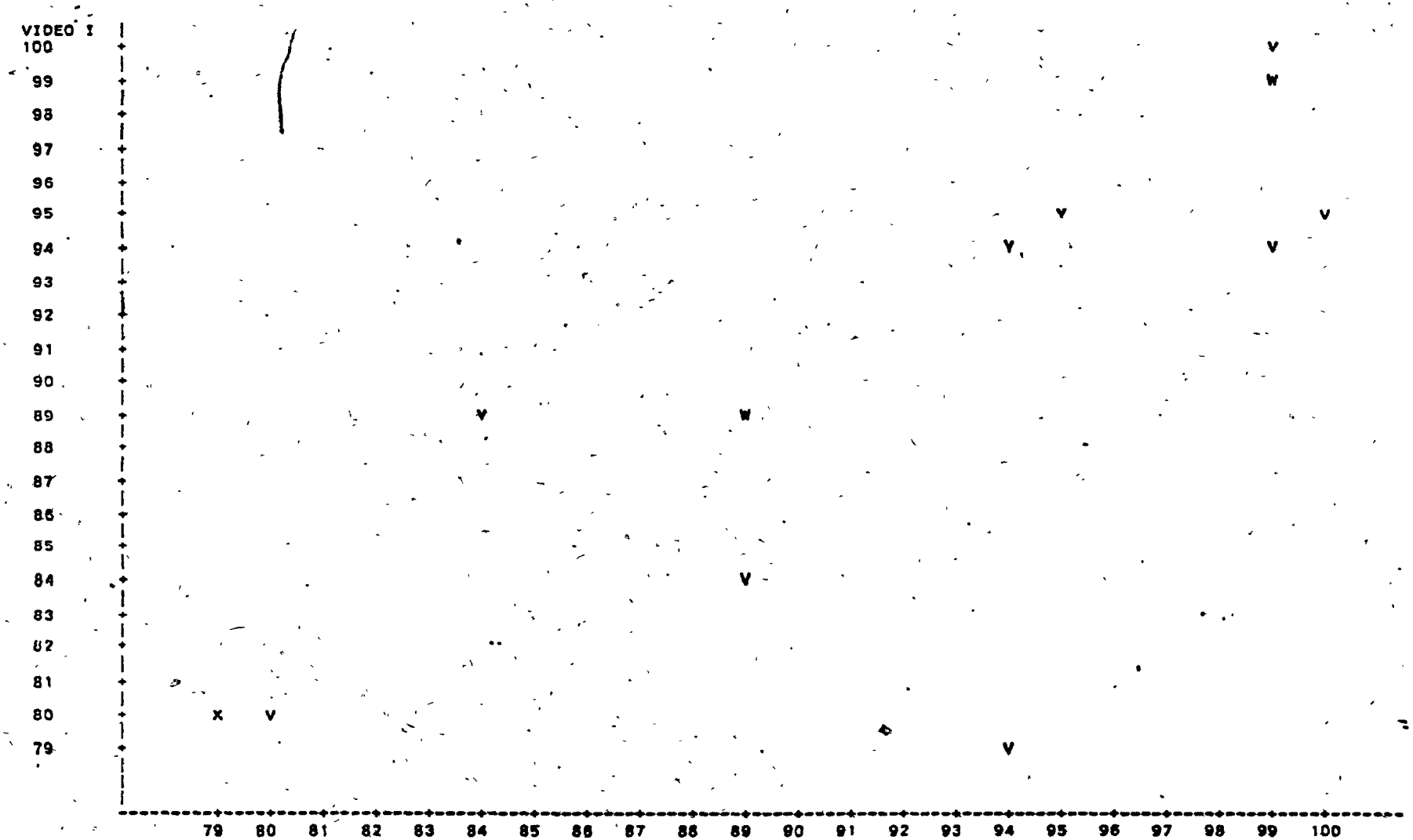
APPENDIX 25

PLOT OF VIDEO I * VIDEO II USING THE BARTHEL INDEX
GROUP II=5 INTERPRETERS (Q TO U)



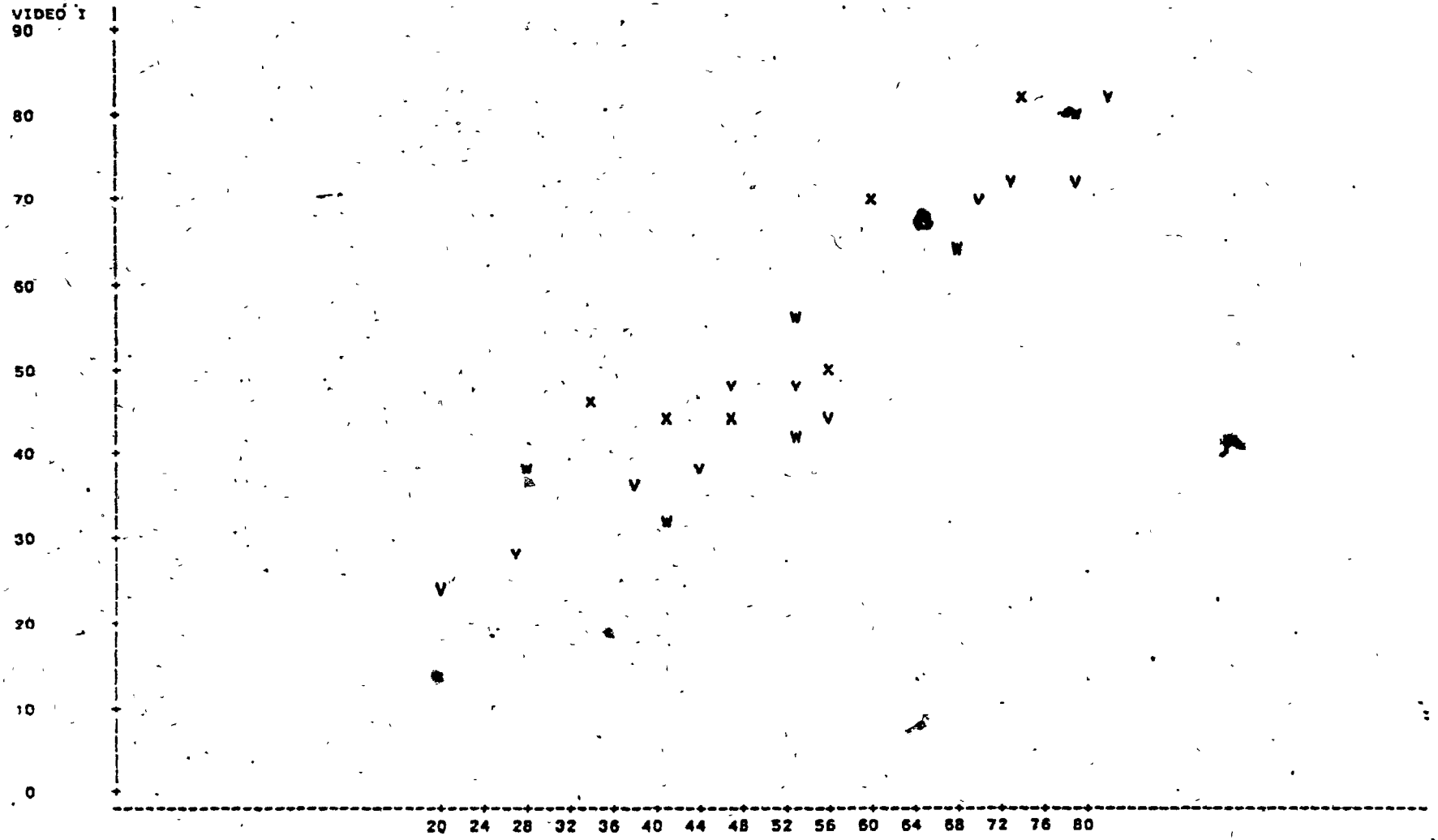
APPENDIX 26

PLOT OF VIDEO I * VIDEO II USING THE LEVEL OF REHABILITATION SCALE
GROUP II=5 INTERPRETERS (Q TO U)



APPENDIX 27

PLOT OF VIDEO I * VIDEO II USING THE BARTHEL SCALE
GROUP III=4 INSTRUCTORS (V TO Y)



PLOT OF VIDEO I * VIDEO II USING THE LEVEL OF REHABILITATION
GROUP III =4 INSTRUCTORS (V TO Y)

APPENDIX 28

APPENDIX 29

Coefficients of Variation % (Intra-Observer) for the Subscales and the Total Score of the Barthel Index and the LORS

GROUP I

	Sources of Variation			
	var(rater in group) +	var(pt) ++	var(r-pt * in group)	var(vid ** + error)
Barthel Index				
Self Care	1.1	2.90	4.3	0.006
Continence	1.5	0.00	4.6	0.008
Mobility	7.5	17.70	9.5	2.400
Total Barthel	2.4	7.10	6.6	0.900
LORS				
Home Activities	7.5	35.0	12.9	2.0
Outside Activities	12.8	45.0	23.0	2.4
Social Activities	6.1	58.0	16.0	5.6
Total LORS	8.4	30.3	9.1	4.2

\+ var(rater in group)=variation of rater in group
++ var(pts)=variation of patients
* var(r-pts in group)=variation of rater by patient in group
** var(video+error)=variation of video session + random error

APPENDIX 30

Coefficients of Variation % (Intra-Observer) for the Subscales and the Total Score of the Barthel Index and the LORS

Group II

	Sources of Variation			
	var(rater + in group)	var(pts) ++	var(r-pt * in group)	var(video ** + error
Barthel Index				
Self Care	1.6	1.00	5.5	1.0
Continence	3.3	3.80	7.9	2.5
Mobility	2.4	17.80	5.0	2.9
Total Barthel	2.5	8.40	2.8	2.5
LORS				
Home Activities	5.3	34.0	9.4	2.9
Outside Activities	2.4	51.6	19.7	5.5
Social Activities	2.7	57.0	10.4	4.5
Total LORS	4.3	30.3	8.7	1.3

\+ var(rater in group)=variation of rater in group
++ var(pts)=variation of patients
* var(r-pts in group)=variation of rater by patient in group
** var(video+error)=variation of video session + random error

APPENDIX 31

Coefficients of Variation % (Intra=Observer) for the Subscales and the Total Score of the Barthel Index and the LORS

Group III

	Sources of Variation			
	var(rater in group) +	var(pts) ++	var(r-pts in group) *	var(video + error) **
Barthel Index				
Self Care	0.0	0.0	0.0	0.0
Continence	0.0	0.0	0.0	0.0
Mobility	0.0	20.0	3.7	5.8
Total Barthel	0.0	8.3	1.5	2.4
LORS				
Home Activities	4.3	33.3	13.8	5.2
Outside Activities	0.0	55.5	12.8	10.6
Social Activities	0.0	62.9	8.9	2.1
Total LORS	3.1	3.9	8.1	1.7

\+ var(rater in group)=variation of rater in group
++ var(pts)=variation of patients
* var(r-pts in group)=variation of rater by patient in group
** var(video+error)=variation of video session + random error