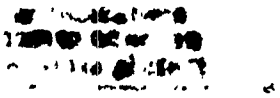
The STroke REhabilitation Assessment of Movement (STREAM):

Content Validity and Preliminary Reliability

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Master of Science in Rehabilitation Science

Katherine A. Daley
School of Physical and Occupational Therapy
McGill University, Montreal
November, 1994

STroke REhabilitation Assessment of Movement: content and reliability.



ABSTRACT

The global objectives of this project were to refine and validate the content, and to obtain preliminary estimates of reliability for the STroke REhabilitation Assessment of Movement (STREAM). Two consensus panels, involving a total of twenty physical therapists, produced an intermediate test version of STREAM. Based on initial evaluations of internal consistency and reliability, items that were redundant, unrelated, and/or not reliably scored were eliminated. The final STREAM, with enhanced content validity, is made up of thirty items evaluating limb movements and basic mobility. Two reliability studies were conducted: 1) a direct observation of twenty stroke patients by pairs of raters in the clinical setting, and 2) two repeated ratings by twenty raters using videotaped assessments of four stroke patients. The STREAM demonstrated excellent internal consistency, inter- and intra-rater reliability, and is now ready for preliminary use in the clinical setting. The favorable results of this study indicate that further testing of the psychometric properties of the STREAM is warranted.

RÉSUMÉ

Ce projet visait le raffinement et la validation du contenu, ainsi que l'obtention de données préliminaires sur la fidélité du STroke REhabilitation Asssessment of Movement (STREAM). La version intermédiaire du STREAM fut développée suite aux suggestions de vingt physiothérapeutes réparties en deux comités indépendents. Suite aux études préliminaires de fidélité et de constance interne, les items redondants, non-reliés entre eux, et/ou non-fidèles furent éliminés. La version finale du STREAM, avec validité accrue du contenu, comporte trente items évaluant la mobilité de base ainsi que la fonction motrice. Deux études de fidélités furent entreprises: 1) l'observation directe par des paires d'évaluateurs dans le milieu clinique de vingt patients ayant subi un ACV, et 2) l'évaluation à deux reprises par vingt évaluateurs des enregistrements vidéo de quatre patients ayant subi un ACV. Dans cette étude, le STREAM a démontré d'excellents résultats tant au niveau de la constance interne que de la fidélité inter- et intra-évaluateurs, et il est présentement prêt pour une utilisation clinique préliminaire. Les résultats favorables obtenus lors de cette étude justifient la poursuite des évalutions psychométriques du STREAM.

PREFACE

In physical rehabilitation following stroke, the overall goal is to improve a patient's functional independence, and thereby improve quality of life. The recovery of functional independence reflects improved motor ability as well as learning of compensatory techniques. Physical therapy treatments in stroke rehabilitation typically endeavor to influence motor ability and learning in order to improve an individual's functioning. That is, therapists apply techniques aimed at enhancing the motor recovery underlying changes in motor abilities. As well, therapists create an environment to facilitate learning by establishing appropriate goals and providing feedback. Although many well established instruments are available for evaluating an individual's overall functional ability, none of the existing clinical tools available for evaluating motor recovery have been widely accepted into clinical practice. It is important, however, that a suitable outcome measure be used routinely for monitoring the motor recovery of individuals undergoing treatment, as this would provide documentation of the rate and extent of motor recovery from hemiplegia. This information is required in order to evaluate the impact that physical therapy interventions have on the recovery of movement in the affected limbs, and thereby to justify physical therapy treatments in the rehabilitation of stroke. Documentation of the recovery of movement in the affected limbs could also further our understanding of how motor recovery contributes to overall functional changes.

A recent survey by the Working Group on Outcome Measures in Physiotherapy (1992), found that several shortcomings, such as lack of clinical utility or unacceptable psychometric properties, have prevented the existing published outcome measures for motor recovery following stroke from being incorporated into routine clinical use. Consequently, our aim was to develop a user friendly instrument that would possess adequate measurement properties, and that would meet the needs of both clinicians and researchers working in stroke rehabilitation.

The STREAM was originally developed and was in use at the Jewish Rehabilitation Hospital in Laval Quebec. This instrument was brief and easy to administer, and provided a convenient means of objectively and quantitatively evaluating many of the limb and basic mobility movements that are routinely subjectively assessed by therapists in the clinical setting. In addition, although no formal evaluations had been made of the instrument's psychometric properties, STREAM showed promise of possessing acceptable measurement properties. Our first step in this project was to carry out a

survey, involving the members of the neuroscience special interest group of the Canadian Physiotherapy Association, to assess the acceptability of the content of the instrument. Although the feedback from the sixty-two survey respondents was generally positive, several possible improvements or refinements were identified. Hence, the objectives of the present study were to improve and validate the content of the STREAM, and to commence the evaluation of the instrument's psychometric properties.

Over the course of this study, the STREAM evolved through several stages. First, the original instrument was refined, based on recommendations made by two consensus panels of experienced therapists, to improve the content validity. Next, this intermediate test version of STREAM underwent preliminary evaluations of item reliability and the relationships between and among items, and item reduction was carried out. This process produced the completed version of STREAM, on which further evaluation of reliability, including the internal consistency, inter- and intra-rater agreement of this completed STREAM, was done. Two separate reliability studies were conducted. A 'direct observation reliability study' was done at the JRH to determine the inter-rater agreement and internal consistency of STREAM, and involved twenty stroke patients and pairs of raters from a group of six participating therapists. And, a 'videotaped assessments reliability study' was done to assess intra- and inter-rater agreement on the scoring of videotaped performances of STREAM, and involved four videotaped assessments that were viewed and rated on two occasions by twenty physical therapists. The five chapters of this thesis present the details of these steps.

In the first chapter, the general need for a clinically useful stoke motor assessment is identified in the significance and rationale for the study. The original version of the STREAM is introduced, and information on how and why it was created is provided. As well, details of a survey on the acceptability of the content of this early STREAM are presented. Next, the conceptual framework, used as a basis for the further development of the STREAM, is established. Finally, justification for the further development and testing of the STREAM is given.

The second chapter is a review of the pertinent literature related to this project. The chapter is organized into three sections. The first section summarizes the motor and functional sequelae of stroke and the recovery of motor function. The topic of the second section is the assessment of impairment and disability following stroke. In this section, issues that complicate the measurement of motor function are identified, and

specific clinical measures that are available for evaluating motor recovery are presented. The third section of this review summarizes the methodology used in developing a clinical scale and in evaluating the psychometric properties of an instrument.

Chapter three presents the study objectives and related methods. This chapter is divided into four sections: the first includes the methods related to refining and validating the content of the STREAM, the second describes the procedures used in the direct observation reliability study, the third describes the methods used in the videotaped assessments reliability study, and the fourth describes the statistical methods used in the analysis.

Chapter four contains the results, which are presented in six sections. First the results relating to the content validity are presented. These consist of: the characteristics of the consensus panel participants, a summary of the evolution of the items and scoring of STREAM, the internal consistency and intra-rater agreement for the items included in the test version of STREAM, and an overview of the process of item reduction leading to the completed STREAM. Subsequent sections present the results from the direct observation and videotaped assessments reliability studies respectively, including sub-sections for the characteristics of subjects and raters, and the rater agreement on items, subscales and on the composite scores of the thirty item STREAM. In addition, the section presenting the results of the direct observation study contains the results of the internal consistency analysis for the completed STREAM.

The final chapter discusses the findings of the study. The chapter includes a discussion of the content development and validation, the internal consistency, and the findings of each of the two reliability studies. In addition, the two reliability studies are compared to one another, as well as to the reliability testing done on related measures. A general comparison is made between the STREAM and related measures in terms of measurement properties and utility. And finally, the limitations of this study and the potential limitations of STREAM as a measure of motor function are identified, implications for future research are discussed, and the conclusions of this study are presented.

At the completion of this project, the revised STREAM with enhanced content validity, and promising preliminary estimates of reliability, is ready for clinical use and for further studies of its measurement properties.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors, Drs. Nancy Mayo, and Sharon Wood-Dauphinee-- two of the most incredibly enthusiastic and productive individuals one could hope to meet. They have been role models "par excellence", and their guidance and encouragement throughout my masters program have been greatly appreciated.

I would also like to express my gratitude to Dr. Irena Danys for her suggestions that helped to establish the theoretical framework in the early stages of this project, and to Roslyn Cabot and Dr. Danys for their helpful feedback on reviewing the various stages of this thesis as members of the thesis committee. Thanks are also due to Kathy Berg, whose study on developing a measure of balance in the elderly served as a model and inspiration for this study. I also wish to acknowledge the patience and assistance of the staff at the Jewish Rehabilitation hospital--their support greatly facilitated the project. A considerable number of therapists participated in the various phases of this project, and I sincerely hope that the end product--the completed STREAM and test manual--will meet their needs and expectations, and will reward them sufficiently for their enthusiasm and cooperation. Special thanks to Lora Salvo for her participation in preparing the videotapes of patient assessments. It goes without saying that this project would not have been possible without the cooperation of the many patients that agreed to participate in the reliability studies--and to them I am deeply indebted. In addition, thanks to Susie Rosenmeier, Sarah Mayo, and Michelle Corrigan for their help in facilitating the consensus panel meetings and videotape viewing sessions as research assistants on the project. Also, special thanks goes to Susan Scott for her invaluable tutelage for the statistical analysis and on the use of SAS, to Robyn Tamblyn for assistance with the GENOVA program, and to Claudette Corrigan for her frequent technical advice and general encouragement. And last, but certainly not least, I would like to thank my family and friends out west, who have not forgotten me through these three long years, and who have continually pressed/encouraged me to "hurry up and get the job done, so you can come home".

This project was funded by a grant from the Réseau de Recherche en Readaptation de Montréal et de l'ouest du Québec. I would also like to acknowledge the financial support I received through a Royal Canadian Legion / Physiotherapy Foundation of Canada fellowship in gerontology, and through the Fonds de la Recherche en Santé du Québec.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

RATIONALE AND BACKGROUND

1.1 Significance and Rationale for the Study

Stroke is one of the major causes of mortality and disability in the population, with approximately 20,000 deaths due to stroke and a further 35,000 individuals afflicted in Canada each year (Statistics Canada, 1989). Survival rates for stroke have increased in the past 25 years, with a concomitant rise in the number of individuals who must live with the sequelae of stroke (Bisch, 1989; Bonita, 1992; Bonita et al, 1988; Kotila, 1984; Mayo et al, 1991a; Mayo, 1993; WHO, 1989). Approximately 75% of survivors of acute stroke exhibit moderate to severe motor and functional impairments initially, and require some type of rehabilitation (Jerntorp et al, 1992; Schmidt et al, 1988); a small proportion (10% to 20 %) of stroke survivors require permanent institutionalization (Mayo et al, 1989; Wade, 1992). Stroke, therefore, is a major social and economic burden, the magnitude of which will continue to increase as survival improves and the proportion of elderly in the population increases (Broderick et al, 1989).

Heightened demand for hospital resources has led to an increased need for evaluating the effectiveness of the clinical management of stroke. Physiotherapists are being challenged to critically analyze current methods for assessing and treating stroke, and to justify treatment programs based on outcomes. A task force, supported by Health and Welfare Canada, has recently developed a strategic plan for encouraging the use of standardized outcome assessments in physiotherapy (Working Group on Outcome Measures in Physiotherapy, 1992). Underlying the goals of this Task Force is the knowledge that reliable and valid outcome measures are required for evaluating the effectiveness of interventions.

In addition to having acceptable reliability and validity, an ideal outcome measure should include sufficient detail to be sensitive to clinical change, and should be wide enough in scope to adequately measure the attributes and populations for which it was intended. As well, for an instrument to be realistically incorporated into the rehabilitation program, it must be easy to administer so that it requires minimal time, equipment, or training, and it has to be acceptable to the subjects being evaluated. The measures that are currently available for assessing outcomes of stroke rehabilitation

vary widely in terms of content and clinical utility, and also in terms of their psychometric properties and the degree to which these have been assessed.

The overall goal of stroke rehabilitation is to minimize the impairments, disabilities and handicaps associated with stroke and, thereby, maximize quality of life. The principal contribution of physical therapy in this process is to improve motor function--specifically, to facilitate the recovery of voluntary movement of the affected limbs and mobility (Carr and Shepherd, 1987; Duncan, 1991). Thus, a measure that is sensitive to these changes in motor function is required for evaluating the impact that physiotherapy interventions have on the rate and quality of recovery from hemiplegia. Measures sensitive to changes in the quality of movement of the affected limbs are also required to enable researchers and therapists to monitor the course and pattern of motor recovery, and for determining the contribution that motor recovery makes to overall functional changes.

Despite the importance of evaluating motor recovery, there was no widely accepted measure available to physical therapists for this purpose. Although several instruments currently available for monitoring the motor recovery of stroke patients have been used for research purposes (Badke and Duncan, 1983; Bernspang et al, 1987; Dettman et al, 1987; Gowland, 1982; Henley et al, 1985; Loewen and Anderson, 1990), they have not been widely employed in clinical practice. A recent Canadian survey revealed that existing published instruments for motor evaluation following stroke are used routinely in only 5% of physiotherapy departments (Working Group on Outcome Measures in Physiotherapy, 1992). Among the reasons cited as barriers to the routine use of existing outcome measures were: 1) lack of knowledge or availability of instruments, 2) poor clinical utility due to complexity of scoring, administration time, and/or dependence on equipment that may limit portability, and 3) a lack of comprehensiveness for assessing the characteristic(s) of interest. A clinical instrument addressing these issues was clearly needed.

1.2 Background of STREAM

In 1986, a scale to measure the recovery of movement following stroke was developed through a collaborative effort between researchers and physical and occupational therapists at the Jewish Rehabilitation Hospital (JRH), Chomedey-Laval, Quebec. This original Stroke Rehabilitation Assessment of Movement (STREAM) (Appendix 1), was intended to provide a relatively comprehensive, standardized, objective and quantitative

assessment to be used by physiotherapists for measuring the motor recovery of stroke patients. The items included in the original instrument were adapted from clinical experience, and from existing published assessments (Bobath, 1978; Brunnstrom, 1970; Fugl-Meyer, 1975). A wide range of movement patterns, typically assessed by physiotherapists for the purpose of evaluating the motor status of stroke patients, were included. The original instrument comprised thirty-four items including fifteen items assessing limb movement patterns (nine for the upper extremity; six for the lower extremity), and nineteen basic mobility items (of which four items were performed in sitting, six in standing and nine involved walking). The items assessing limb movement patterns, and mobility items performed in sitting, were scored on a dichotomous scale (0:unable; 1:able to perform the test item); mobility items performed in standing and walking were scored on a four-point ordinal scale related to degree of assistance required (0:unable; 1:requiring the help of another individual; 2:with the help of an assistive device; 3:able to perform the test item independently and safely with no aids). The maximum total score on the STREAM was sixty-four. The scale was piloted for clinical utility at the JRH for a number of years and proved to be clinically acceptable, but no formal evaluation of the STREAM had been made prior to this study.

1.3 Content Verification Survey

As a preliminary step in this study, a content verification survey (Appendix 2) was carried out to assess the broad acceptability of the original STREAM. A copy of STREAM, along with a stamped return envelope, was sent to a sample of convenience of 310 physiotherapists involved in the management of neurological conditions. The majority of those sampled were members of the Neuroscience Division of the Canadian Physiotherapy Association. Sixty-two therapists responded to the survey [1].

Therapists' opinions were solicited regarding: 1) the importance of each item; 2) the clarity and appropriateness of the scoring; 3) the adequacy of the STREAM as a measure of motor function; and 4) the extent to which the respondents would use STREAM in their clinical practice.

---

1 Although only 20% of those surveyed responded, health professionals do not usually respond to mailed surveys without extensive encouragement and follow up (Arsenault and Cleather, 1982). This survey was not designed to estimate any parameter or test any hypothesis, rather its aim was to obtain comments from interested persons as to the utility of STREAM.

The therapists were asked to rate the importance of each item included in the STREAM using a five-point scale (1: crucial; 2: very important; 3: moderately important; 4: neither important nor unimportant; 5: unimportant). The criteria set for retaining items required that at least 80% of the ratings given by panel members had to be in the range from 1 to 3, and less than 5% of the ratings could be in category 5.

Based on the above criteria, only three of the thirty-four items were rejected: four of the sixty-two respondents (6.5%) rated items 2, 5 and 18 as unimportant. The reasons given for rejecting these items related to the potential for interference by shoulder pain and the non-functional orientation of the items.

Table 1.1 indicates the responses made to the survey questions. Overall, the responses to the survey questions were favorable, except for the question of whether the STREAM would be adequate to assess motor function. The majority commented that, if used in isolation, the STREAM would not be adequate to plan treatment. This was a point of confusion, however, as the scale was not intended for planning treatment strategies, but rather for evaluating treatment outcomes.

TABLE 1.1
Results of the Content Verification Survey
(62 persons responding)

Survey Question	Proportion Agreeing
STREAM adequate to assess motor function	51%
Items clear	74%
Scoring appropriate	74%
Adequate for showing differences in patient status	89%
Would use instrument	70%

Despite generally positive feedback, the survey identified many possible refinements which could be made to improve the instrument. The most commonly expressed suggestions were that the STREAM should include: more details relating to the quality of movement; more details on postural/trunk control; more mobility items; and more lower level limb movement items. In addition, it was proposed that the original

scoring (0: unable; 1: able to perform the test item) should be modified to increase the response options in an effort to improve the sensitivity to change over time. Finally, it was apparent that there was little consensus regarding the domains that should be included in a clinical measure of motor function.

The information from the survey influenced the further development of the study protocol. Rather than commencing with clinical reliability and validity studies as had initially been proposed, a more preliminary level of scale development was planned. Thus, to enhance content validity, further work on the content of STREAM was initiated. The comments of the sixty two survey respondents were used as a starting point for developing the conceptual framework described below, and for revising the instrument as indicated in the methods section for content validation (3.1.2).

1.4 Theoretical Framework

Given the complexity of motor performance in terms of the number of factors that potentially influence motor recovery and functional performance following a stroke, defining this phenomenon in operational terms presented a challenge. A variety of terms could have been used to describe what the STREAM endeavors to measure. Included are motor recovery, motor ability, motor function, functional motor recovery, motor performance, motor control, motricity, and motor output. Unfortunately, many of these terms are not well defined and may imply different concepts to different individuals. We selected the terms "voluntary movement" and "basic mobility" to describe the attributes measured by the STREAM, and the following conceptual framework was developed.

Motor recovery implicates a number of neurophysiologic processes at the cellular level. In addition, factors such as the size and side of the lesion, age, comorbid conditions, motivation, communication, cognition, contractures and pain impact on recovery. Although it is very difficult to separate the relative influences of these various processes and factors, motor recovery is manifested by the re-emergence of voluntary movement and restoration of basic mobility. Therefore, the STREAM is intended to measure, in the clinical setting, these fundamental building blocks that reflect motor recovery.

The World Health Organization's International Classification of Impairments, Disabilities, and Handicaps (ICIDH) (WHO, 1980) has been widely used as a framework within which treatment goals and outcomes of stroke rehabilitation are

described (Duncan, 1991; Granger and Gresham, 1990; Task Force on Stroke, 1990). Within this model *impairments* refer to the primary deficits of anatomical, biological, psychological or physiological structure or function observed as a result of disease or injury. *Disability* refers to the functional consequences of these impairments. Thus, in the context of physical rehabilitation following stroke, impairments such as paresis, sensory deficits and abnormal tone may contribute to, or cause, motor disability such as reduced ability to perform purposeful movements, transfers, walking, and activities of daily living (ADL) (Granger, 1984; Grimby et al, 1988; Guccione, 1990). In this framework, the STREAM is related hierarchically to other measures of impairments and disabilities in that it is intended to measure basic motor ability--one step beyond the level of the primary impairments, and one step before functional mobility and ADL measures.

1.5 Rationale for Further Developing and Testing the STREAM

The STREAM may satisfy the measurement needs of physical therapists. Hopefully, it will be sufficiently comprehensive, yet concise enough, to be attractive for use in both clinical and research settings. Contingent on the acceptability of the measurement properties and clinical utility of the revised STREAM, this instrument could ultimately streamline the process of clinical motor evaluation, improve the communication between and within clinical and research settings, and provide an expanded data base for future research in stroke rehabilitation. Further refinement, and evaluation of the psychometric properties of STREAM was therefore warranted to enhance this scale's potential for use in measuring motor recovery following stroke.

CHAPTER 2

LITERATURE REVIEW

2.0 Overview of the Literature Review

The literature review focuses on the following topics: the clinical sequelae of stroke and recovery of motor function, the assessment of impairment and disability following stroke, the stroke motor function assessments currently available for use in the clinical setting and in research, and the basic requirements of scale development. The first section of the review summarizes the motor and functional changes, and the natural progression of motor recovery typically observed following stroke. Next, a number of general and stroke specific measurement issues and controversies are noted. The methods currently in use for evaluating motor recovery are outlined, with the main focus of this section being on clinical indices of motor function. Finally, the methodology involved in developing and testing clinical indices is presented in detail. The information provided in each of the topical areas of this review was used in developing the study protocol, and in carrying out the project. Also, throughout the literature review the significance and rationale for the study, presented in the previous chapter, are further substantiated.

2.1 Motor and Functional Sequelae Following Stroke

2.1.1 Clinical presentation of stroke

Cerebral vascular accident (CVA), or stroke, is a rapidly developing neurologic dysfunction due to a disturbance of cerebral circulation, with symptoms persisting for more than twenty four hours or leading to death (World Health Organization (WHO), 1989). Stroke may be due to infarct or hemorrhage in the distribution of the cerebral vessels, with the majority of cases (80 to 95%) being due to cerebral infarct (Bogousslavsky, 1988). The pathology of stroke varies widely in terms of cause, extent and location of vascular disturbance, yet a considerable degree of similarity of neurological signs is observed among patients.

The classic sign is a unilateral motor deficit, of varying severity. Hemiparesis, or weakness on one side of the body, is displayed by 70 to 80% of stroke patients (Walker et al, 1981; WHO, 1989). Other, often concurrent, motor deficits include ataxia of limbs and/or trunk and apparent motor deficits resulting secondary to sensory disturbances. Changes in reflexes and muscle tone frequently accompany motor

deficits, typically presenting as early flaccidity and diminished reflexes with gradually increasing spasticity over the ensuing days and months (Brunnstrom, 1970; Fugl-Meyer et al, 1975; Twitchell, 1951). Communication, continence and cognition are frequently affected to varying degrees and can further impact on motor function.

2.1.2 Recovery of voluntary movement

Twitchell (1951) followed the course of recovery of 19 patients and described a typical pattern of motor recovery following stroke, that varied with respect to timing and extent of recovery. Recovery of movement generally began proximally, and gradually more distal segments regained motor power. Early movement was typically limited to a synergistic pattern of total extension for the lower extremity and total flexion for the upper extremity, although not all subjects fit this pattern precisely. Voluntary movement control gradually emerged so that movement out of synergy was possible. Others (Brunnstrom, 1970; Fugl-Meyer et al, 1975) have documented a similar pattern of motor recovery.

More recently, patterns of recovery of muscle strength (as well as tone and reflexes) have been studied (Bohannon, 1988a; Demeurisse et al, 1980; Gray et al, 1990; Wade et al, 1985; Wade and Hewer, 1987). Although these authors have described motor recovery from a different perspective, their findings do not contradict Twitchell's early observations. However, the neurophysiological basis of the synergistic patterns of recovery, previously considered to be a result of recovery of a hierarchically organized neural control system, has been questioned. Biomechanical factors, such as imbalance of strength in opposing muscles, the effects of gravity, and muscle shortening or contracture, have been implicated for explaining the synergistic patterns typically observed (Carey and Burghardt, 1993; Corcos, 1991; Craik, 1991; Dietz, 1992).

In general, the extent of recovery of movement in the lower extremity has been found to be greater than that observed in the upper extremity (Gowland, 1982; Partridge, et al, 1987; Wade and Hewer, 1987). Approximately 1/3 of persons suffering a stroke make no functional recovery in the upper limb, while 1/3 make partial recovery sufficient to use the arm for some functional activities, and the remaining 1/3 make more complete recovery. For the lower extremity, only 10% have severe permanent motor loss (Wade and Hewer, 1987). A more recent study (Duncan et al, 1994), however, has reported that the recovery from motor impairments in upper and lower limbs was equivalent. The authors suggest that neurological recovery in the limbs is parallel, but because upper extremity tasks typically involve finer and more complex

control than lower extremity tasks, the limbs may differ in terms of functional ability. The majority of patients will make most of their motor recovery during the first three months (Duncan et al, 1994; Kelly-Hayes et al, 1989; Wade et al, 1985). The initial severity of the neurologic deficit and the rate of early recovery of movement are strongly predictive of the eventual degree of motor recovery (Dove, 1984; Jongbloed, 1988; WHO, 1989).

2.1.3 Recovery of functional ability

Recovery of function may continue well beyond the early dramatic period of motor recovery, particularly in terms of adaptation and independence in functional ability (Andrews et al, 1981; Mayo et al, 1991b; WHO, 1989). In addition to the initial severity of motor impairments, the rate and degree of functional recovery have been found to be influenced by the following factors: location of the cerebral lesion, age, presence and severity of perceptual deficits, comprehension difficulties, depression, history of prior stroke, coexisting medical conditions, sensory deficits, loss of sitting balance, and incontinence (Bonita and Beaglehole, 1988; Dove et al, 1984; Gowland, 1982; Jongbloed and Jones, 1988; Loewen and Anderson, 1990; Mayo, 1991b; Olson, 1990; Osberg et al, 1988; Sandin and Smith, 1990; Shah et al, 1989; WHO, 1989). Thus, the motor, sensory, cognitive, and emotional sequelae of stroke are all potential contributors to functional disability. Training and motivation may also impact on motor and functional recovery (Duncan and Badke, 1987).

In terms of independence in ADL, 1/2 to 2/3 of survivors achieve independence or have only a slight dependency while the remainder require moderate to total assistance (Bonita, 1988; Wade and Hewer, 1987). The ability to walk alone, with or without assistive devices, is achieved by 50% to 85%of survivors (Anderson et al, 1979; Gresham et al, 1979; Moskowitz et al, 1972; Wade et al, 1987). Hospitalization and functional training are required by the majority of stroke survivors in order for maximal functional levels to be achieved (Gresham et al, 1975; WHO, 1989).

2.2 Measurement of Motor Performance Following Stroke

2.2.1 Issues in the measurement of motor function

The World Health Organization's International Classification of Impairments, Disabilities, and Handicaps (ICIDH) (WHO, 1980), has become the standard framework for describing the consequences of disease or injury. Moreover, the terminology of this classification system is commonly used for defining the focus of

physical therapy assessments and treatments. In physical therapy for stroke rehabilitation, the evaluation of disability involves establishing if the patient can or cannot perform certain activities; the evaluation of impairment focuses on the factors contributing to, or causing the dysfunction (Granger, 1984). For example, instruments that measure activities of daily living (ADL), such as the Barthel Index (Mahoney and Barthel, 1965), are classified as disability measures, and instruments that measure sensation, spasticity or strength, are impairment measures.

The distinction between impairment and disability measures becomes less clear when one considers motor function (Jette, 1985; Granger, 1984). Motor function involves the integration of many discrete components, such as: sensation, range of motion, pain, coordination, tone, strength, vestibular function, and cognition (Duncan and Badke, 1987). Impairment of any of these components can effect motor function. Some instruments measuring motor function attempt to discriminate between and identify the contributions made by the individual components of motor function, and include subscales that evaluate pain, range of motion, and sensation, along with those evaluating limb movement, balance and mobility (Fugl-Meyer et al, 1975; Gowland et al, 1991). Under the ICIDH framework, subscales measuring pain, tone, and strength fall into the category of impairment measures, while mobility subscales are classified as disability measures (WHO, 1980). Subscales evaluating limb movements are generally classified as motor impairment measures (Gowland et al, 1991, 1992, and 1993; Huijbregts, 1992; Sanford et al, 1993; WHO, 1980). However, limb movements included in motor function evaluations often tend to be related to functional tasks, and to reflect the integration of strength, tone, range, sensation, and/or coordination. A recently proposed modification of the ICIDH (Jette et al, 1994) incorporates an intermediate category between impairments and diabilities. This category, "functional limitations", includes any restriction or lack of ability to perform an action in an expected manner or range. Thus, disturbances of voluntary movement of the limbs would be included in this additional category along with basic mobility and ambulation. This schema may be more suitable for classifying instruments that are oriented towards measuring functional motor recovery rather than measuring pure motor impairment.

Similarly, there is a lack of a clear distinction between measures of motor function and measures of functional ability. While functional ability typically refers to independence in ADL, involving elements of cognitive and perceptual functioning, motor function relates to more basic movements (Granger, 1984). A grey area or overlap exists

between these two types of measures, and activities such as moving between lying and sitting or standing, walking, or upper extremity tasks such as reaching and grasping may be included in both functional independence and motor function evaluations. As the ability to integrate movements in the performance of motor tasks is an important aspect of motor recovery, including items that involve basic mobility, upper extremity movements across more than one joint, or simple goal oriented tasks in measures of motor function is justifiable.

Because the recovery of functional independence is a major goal of rehabilitation, and indices measuring the performance of ADL are frequently used as measures of rehabilitation outcome, they are also sometimes used to reflect changes in the motor status of individuals who have had a stroke. Functional independence measures, however, do not specifically test the motor recovery of the affected limbs. In the absence of motor recovery much improvement in the performance of ADL can be accomplished by learning compensatory techniques (Gowland, 1982; Jette, 1984). Functional independence measures, thus, summarize the complex integration of cognitive, affective and sensorimotor abilities, and, therefore, are of limited use for reflecting pure motor recovery.

Another issue relates to the distinction between the terms "measurement" and "assessment". Although in practice these terms are often used interchangeably, they are not necessarily synonymous. "Measurement" relates to the standardized procedures used to quantify the extent or the quality of an attribute or characteristic, while "assessment" infers a more comprehensive evaluation, often involving a more "qualitative" or non-standardized approach, and interpretation of findings (Craik and Oatis, 1985; Leahy, 1991; Wade, 1992). Detailed assessments are required for the purpose of planning treatments. For overall evaluation of a treatment approach, or for operationally defining a characteristic in order to facilitate comparison between individuals, the objective is to obtain a quantified summary of the characteristic(s) under study, and a detailed evaluation may not be required or practical.

Because the intended use of the measure dictates the amount of detail included and the characteristics targeted, several closely related guidelines have been used for defining the purposes of measurements (Feinstein, 1987; Kane and Kane, 1981; Kirshner and Guyatt, 1985). In general, these schemes include the following categories of measurement purposes: 1) describing patients, 2) predicting recovery, and 3) evaluating the effects of treatment. In stroke rehabilitation, for example, a therapist

might quantify characteristics of individual patients, and use this information to compare patients, predict the probability of their full recovery from stroke, or assess the effect of a treatment approach. For each of these measurement purposes, the content and psychometric requirements may differ in some respects (Kirshner and Guyatt, 1985). An instrument intended for describing individuals within groups should be comprehensive, and measure all aspects of attributes that would discriminate between individuals; the information provided by such measures can be used to identify and assess the extent of the specific disabilities impacting on an individual's performance, and subsequently lead to the appropriate choice of therapeutic interventions. In contrast, predictive instruments are generally used for screening individuals for disease or risk factors; these instruments, therefore, must be quick to administer, and accurate, but are not required to be detailed or comprehensive. Evaluative instruments, used for monitoring the clinical status of patients and for evaluating the effectiveness of treatments, must be sensitive to clinical changes over time; they focus on measuring characteristics that are expected to change, therefore, tend to be less extensive in scope but include more sensitive and select details than descriptive measures (Kirshner and Guyatt, 1985). Ideally, instruments could be developed for serving multiple measurement purposes. Although achieving such versatility may require some compromises in terms of meeting the somewhat conflicting criteria outlined above, the possibility of using the same instruments for several purposes is appealing.

In addition to the conceptual issues and controversies described above, there are many methodological choices to be made in developing or selecting instruments to measure motor function. Once the conceptual framework and purpose of the measurement are defined, appropriate and valid methods for measuring the phenomenon of motor function must then be selected. For example, one approach would be to consider motor function in terms of component parts or domains, such as limb movements, postural stability (trunk control), and mobility, and then to evaluate a patient's performance on standardized activities of graded difficulty within each of these domains. Moreover, numerous parameters may be used to reflect motor function, including: the degree of assistance or help required, the time taken to perform a task, the degree of difficulty perceived in performing the activity, and the pattern used or proficiency of the movement (Gans et al, 1988). Deciding between measuring the ability to complete a task versus measuring the quality of the performance can present a dilemma; measuring both aspects, if feasible, could be the best solution (Craik, 1991).

A related problem is whether to measure usual performance, or performance under favorable circumstances. Again, the approach taken will depend on the intended use of the measure; for prediction of performance after discharge the former approach may be selected, while for program planning the latter may be the more desirable strategy (Keith, 1984). Yet another choice is whether to examine the isolated motor function of the involved side exclusively (possibly including some aspect of balance or mobility, sensation, range of motion, or pain-- as do many motor function measures for stroke), or to include measurement of the so called "uninvolved side", which in fact is frequently found to be involved to some extent following stroke (Craik, 1991; Lindmark and Hamrin, 1988). Finally, decisions must also be made regarding mode of administration, and regarding required qualifications and training of the evaluator. Observer ratings of actual performance are more reliable and valid (albeit more time consuming and expensive, particularly when trained professionals are required to administer the test) than are self-report measures, especially for use with cognitively impaired stroke patients (Bergner, 1987).

In summary, an ideal approach to measuring motor recovery is yet to be established, and a good deal of confusion surrounds the issue of how to define motor function, let alone how to assess this construct optimally. Using the ICIDH or similar classification system to describe the level of disease that is the target of a measure, and defining the purpose of the measurement in terms of the guidelines described by Kirshner and Guyatt (1985), should resolve some of the issues noted above. Consistent use of the standardized terminology of these classification schemes would improve interdisciplinary communication, facilitate the appropriate use of instruments, and minimize confusion related to instrument construction and use.

2.2.2 Laboratory measures of motor function

Neurophysiologic studies endeavor to identify the specific motor pathways disturbed and to understand the impairments underlying the disability. Laboratory tests to assess motor function might include muscle fiber biopsy, measurements of torque generated during controlled tasks, and/or electromyographic and kinematic analysis of patterns of activity in selected groups of muscles during controlled voluntary movement or functional activities (Corcos, 1991; Knuttson, 1979; Knuttson and Martensson, 1980; Rosecrance and Giuliani, 1990; Sjostrom and Fugl-Meyer, 1981). Isokinetic testing has been suggested as a means to evaluate common motor control deficits following stroke, including the ability to generate force quickly, to maintain the force output, and

to terminate force appropriately (Watkins et al, 1984). However, if hyperactive stretch reflexes of antagonist muscles interfere with reciprocal movements, the isokinetic torque measure then becomes an indirect measure of reflex activity and must be interpreted with this in mind. Two instruments, the hand-held dynamometer (Bohannon, 1989; Riddle, 1989) and the Cybex II (Tripp et al, 1991), have been shown to be acceptably reliable for measuring isometric and isokinetic strength of spastic patients. No one laboratory test exists, however, that adequately measures overall motor function or recovery of the affected limbs. Segment by segment analysis of strength or EMG to evaluate the entire system under controlled conditions would be extremely laborious and impractical as a standard means of assessment, and such artificial testing conditions may not reflect actual clinical ability.

Due to the breadth of the domain of motor performance, it would be ideal to combine the results of laboratory measures such as neurophysiologic and kinematic tests, clinical measures of impairments such as strength, sensation, and ROM assessments, and standardized clinical motor function evaluations. This would provide a more detailed profile of a patient's motor function, and a link between neurophysiologic mechanisms and clinically observed changes in motor function (ie. a link between impairments and disability).

2.2.3 Clinical indices for evaluating motor dysfunction following stroke

Numerous clinical scales have been developed to measure motor function following stroke. These indices vary widely in terms of conceptual basis, domains measured, population targeted, mode of administration and scoring, to name just a few differences.

Instruments have been developed for measuring isolated motor skills, such as walking (Gronley and Perry, 1984; Holden et al, 1986), balance (Berg, 1989; Horak, 1987; Shumway-Cook and Horak, 1986), mobility (Jebsen et al, 1970; Seaby, 1987 & 1989; Tinetti, 1986) or hand function (Carroll, 1965; DeSouza et al, 1980; Jebsen et al, 1969; Lyle, 1981). Brief screening tools, such as the Timed Up and Go (Podsiadlo and Richardson, 1991) provide a quantitative summary score for overall basic mobility. There are also instruments available that are not disease specific, such as the Tufts Assessment of Motor Performance (Gans et al, 1988). As well, there are instruments that include the evaluation of cognitive functioning and functional independence in addition to measuring motor function, such as the Burke Stroke Time Oriented Profile (Feigenson et al, 1979).

The detailed qualitative stroke motor assessments developed by Bobath (1978) and Brunnstrom (1970) appear to be the foundation from which numerous quantitative instruments have been devised. These include the Fugl-Meyer Sensorimotor Assessment for Stroke (Fugl-Meyer et al, 1975), the Chedoke-McMaster Stroke Assessment (Gowland et al, 1992), the Motor Capacity Assessment (Lindmark and Hamrin, 1988), the LaVigne Motor Recovery Assessment Scale (LaVigne, 1974), the Rivermead Assessment of Motor Function in Stroke Patients (Lincoln and Leadbitter, 1979), the Evaluation of the Hemiplegic Subject Based on the Bobath Approach (Guarna et al, 1988), and the Physical Assessment for Stroke Patients (Ashburn, 1982).

More recently, a functional task oriented Motor Assessment Scale was developed (Carr et al, 1985). This instrument is based on the motor control model, which assumes that motor function is dependent on the interactions between the central nervous system, musculoskeletal system and biomechanics, and that verbal and sensory feedback, motivation, and practice can influence recovery (Brooks, 1986; Carr and Shepherd, 1986 and 1987; Duncan and Badke, 1987; Winstein, 1987).

An overview of the general characteristics of each of these stroke motor assessments is given in Table 2.1. Of the measures noted, the more frequently used in clinical settings and in research (at least in Canada) include: the Fugl-Meyer Sensorimotor Assessment, the Chedoke-McMaster Stroke Assessment, and the Motor Assessment Scale (Working Group on Outcome Measures in Physiotherapy, 1992). The measurement properties of these instruments have been more extensively evaluated and documented than any of the other instruments, and are reviewed below.

The **Fugl-Meyer Sensorimotor Assessment for Stroke** is a quantitative motor assessment based on Twitchell and Brunnstrom's hierarchical model, which assumes that motor function improves after stroke in a predictable sequence of synergistic patterns. The Fugl-Meyer Scale includes five domains: upper extremity, lower extremity, balancing ability, sensation, and range of motion. A total of 113 items are included, and are scored using a three point ordinal rating (0:cannot perform; 1:performs partially; 2:performs fully). Item scores are summed to give individual domain scores (with upper and lower extremity subscales summing to a maximum score of 100) and a maximum total score of 226. One reliability study involved nineteen chronic stroke patients, and four physical therapists who performed three assessments at three-week intervals. Inter- and intra-rater reliability coefficients were reported to be greater than 0.85 (Pearson's r) for both upper and lower extremity domain subscores

and total score (Duncan et al, 1983). Another study, using three physical therapists, and twelve patients who were less than six months post stroke and undergoing active rehabilitation, reported Intra-class Correlation Coefficients (ICCs) for inter-rater reliability of 0.96 for total score, and 0.97 and 0.92 for upper extremity and lower extremity subscale scores respectively (Sanford et al, 1993). Concurrent validity for the total scale, and predictive validity for the lower extremity subscale have also been demonstrated (Clarke et al, 1983; Dettmann, 1987; Kusoffsky et al, 1982). An analysis of Fugl-Meyer and Barthel Index scores for 167 hospitalized stroke patients, at admission and five weeks later, revealed that Fugl-Meyer and Barthel Index scores were highly correlated (Pearson's r). However, in terms of statistical power for detecting clinical changes, the Fugl-Meyer Assessment was a less efficient outcome measure than was the Barthel Index (Wood-Dauphinee et al, 1990).

Validity of the sequential pattern of recovery has been supported by Fugl-Meyer et al (1975) and others (Brunnstrom, 1970; Twitchell, 1951) in longitudinal studies. Others however have questioned the "synergy" approach, as not all patients follow this rigid sequence of recovery (Brosseau and Potvin, 1993; Carey and Burghardt, 1993; Carr and Shepherd, 1985; Corcos, 1991; Dietz, 1992). Another basic criticism of the Fugl-Meyer Assessment is that it is not functionally oriented, and thus does not conform to the currently popular motor control theory which advocates goal oriented assessment. Also, this scale does not assess gross mobility or ability to ambulate. These higher level motor skills require sequential coordination of body parts. They reflect the ability to integrate movements into functional patterns and are, therefore, important components of motor recovery. The upper and lower extremity subscales have proven to be useful and sensitive to change, however, the balance section is coarse and the sensation and range of motion sections are cumbersome and less reliably measured. In addition, no standardized administration guidelines have been published. These are needed to ensure that testing procedures are consistent across individuals and across institutions. And finally, a major factor which limits the clinical usefulness of this scale is the overall length, as 30-50 minutes is required to administer the full assessment.

The Fugl-Meyer Sensorimotor Assessment was recently modified by Lindmark and Hamrin (1988) to produce the **Motor Capacity Assessment**. This instrument is a composite index of sixty-three items; it includes a mobility component, and measures the motor function of the non-paretic side as well. This modified scale was evaluated

for construct validity, internal consistency, and concurrent validity (comparison with Fugl-Meyer scores) on a large sample of stroke rehabilitation patients (N=231), and it too was reported to be a valid and reliable measure. However, this instrument has the same inherent limitations as the original scale.

Carr and colleagues (1985), developed a functionally oriented **Motor Assessment Scale** for stroke (MAS). Instead of using flexor and extensor movement patterns or synergies, this instrument measures performance of relevant everyday motor activities, and assumes that there is a specific sequence of actions necessary to accomplish each motor task. Eight areas of motor performance pertaining to general mobility, arm and hand function, and one item related to tone, are assessed and scored on a seven-point ordinal scale related to quality of task performance or severity of tone. Some of the items are subjectively described rather than using clearly defined quantitative ranges, therefore, observer rating may be open to bias or ambiguity. Nonetheless, Carr and colleagues (1985) reported high average correlations for inter-rater (r=0.95) and test-retest (r=0.98) reliability on a sample of fourteen stable stroke patients, using twenty raters. Pearson's correlation coefficient, however, does not take systematic bias into account. Although the high correlation indicates that the scoring was parallel, it does not rule out the possibility of some raters consistently scoring higher or lower than the others. More information regarding actual concordance between scores would have been gained had they used ICCs. The developers also studied inter-rater reliability using five videotaped patient assessments and twenty trained raters, with raters' scores compared to a criterion score generated by the author. Average percent agreemen between scores was reported to be 87%. Crude agreement, however, does not correct for the amount of agreement expected to occur by chance. Poole and Whitney (1988) assessed the interrater reliability of the MAS using direct observation of assessments of twenty-four patients with a wide range of motor deficits rated by two raters. They reported a correlation (Spearman's rho) of 0.99 between total scores, with all items highly correlated except for tone.

The MAS was modified by Loewen and Anderson (1988) by dropping the item for tone, which they felt was too subjective. They used seven videotaped assessments and fourteen raters, and reported inter-rater and intra-rater reliability of the modified MAS with excellent agreement for 80% to 85% of scores based on calculated chance corrected Kappa statistics.

The validation of the MAS is as yet in the preliminary stages. Concurrent validity was assessed by correlating MAS scores with Fugl-Meyer scores on paired items (Poole and Whitney, 1988); all scores were highly correlated except for sitting balance, possibly indicating that the MAS item reflects dynamic rather than static balance. Information on responsiveness has not yet appeared in the literature. While this scale has been met with enthusiasm because of its non-synergic, functionally oriented approach to measuring motor performance, it is not a comprehensive assessment of motor function in that it lacks details of limb movement patterns for the lower extremity. These have been subsumed in the mobility and walking items. Thus, the MAS yields a relatively general summary of motor function, and provides less insight into the specific areas of the body where recovery is occurring than do other measures that are organized into more detailed subscales for upper and lower limb motor function and mobility. Another inconvenience in using this scale in clinical practice is that a number of props, such as jelly beans, pen caps, and tea cups, are required for administering the assessment.

The **Chedoke-McMaster Stroke Assessment** (C-McSA) is a recently developed clinical assessment which is currently undergoing evaluation of its measurement properties. It is intended to be a comprehensive clinical measure for evaluative, predictive and discriminative purposes. Conceptually the scale is based on the WHO's ICIDH, and utilizes Brunnstrom's synergies. The instrument includes a physical 'impairment' inventory (six items measuring the stage of recovery of the arm, hand, leg, and foot, postural control, and shoulder pain) and a disability inventory (fourteen items measuring gross motor function and one item assessing walking ability). Each item is scored on a seven point ordinal scale (related to stage of recovery for the impairment inventory and to degree of independence for the disability inventory) except for the two minute walking test which is worth two bonus points if performance is normal. The maximum score for the impairment inventory is forty-two; the maximum score for the disability inventory is 100.

A study was carried out to determine inter-rater, intra-rater, and test-retest reliability, the concurrent and construct validity of the entire measure, and responsiveness of the mobility component (Gowland et al, 1992 and 1993). Inter-rater agreement was assessed by having two therapists score a sample of thirty-two in-patient and day hospital stroke patients during their first week following admission. The physical impairment inventory had ICCs ranging from 0.85 to 0.96, with an ICC for the total

score of 0.97. The ICCs for the disability index were 0.98 for the individual items and 0.99 for the total score. Intra-rater agreement was assessed using admission scores, and videotaped admission assessments of the same stroke patients scored by a therapist after a minimum interval of two weeks. The ICCs for the physical inventory items ranged from 0.93 to 0.98 and was 0.98 for the total impairment inventory.

Construct validity was assessed by comparing specific items on the C-McSA with selected similar components of the Fugl-Meyer Sensorimotor Assessment and Functional Independence Measure (FIM). All correlations were found to be high, with the exception of a moderate level of association found between shoulder pain and upper limb pain scores (r=0.76). Concurrent validity was assessed by comparing the C-McSA impairment inventory scores with total Fugl-Meyer scores (r=0.95), and disability inventory scores with FIM scores (r=0.79). In addition, variance ratios for admission to discharge change scores (obtained by dividing the variance due to change by the sum of the variance due to change and error variance, with values close to one reflecting maximum responsiveness) (Norman, 1989) were 0.53 for the disability inventory as compared to 0.39 for the FIM, suggesting that the C-McSA disability inventory is more responsive to change than the FIM.

The theoretical basis of the C-McSA has been presented in detail (Gowland et al, 1991 and 1992; Huijbregts, 1992; Moreland et al, 1993). The content validity of the disability inventory has been reported; thirty-one patients and twenty-seven care givers were asked to rate the items on a seven point scale in terms of perceived importance (1:unimportant to 7:very important), and the mean score given for all items was above six (Huijbregts, 1992). The content validity of the 'impairment' inventory, however, has not yet been reported. The overall assessment of reliability and validity, although preliminary, has thus far been exemplary and the results support the use of this scale. As with the Fugl-Meyer motor assessment, however, a dependence on synergistic patterns of movement and the overall length of the assessment are two inherent drawbacks.

Table 2.1 summarizes the available information relating to the general characteristics and psychometric properties of the three measures presented in detail above, and for the majority of published stroke motor assessments. As can be seen in the table, the information available on many of these instruments is extremely limited. They have, however, been included for completeness, and because the information that was available on these instruments was helpful in the present project.

TABLE 2 . 1

Overview of Characteristics of Stroke Motor Assessments

Scale	Content Domains	Conceptual Framework	Scoring	Administration Time	PSYCHOMETRIC PROPERTIES						
					Reliability			Validity			Responsiveness
					Inter-rater	Intra-rater	Internal Consistency	Content	Criterion	Construct	to Clinical Change
Fugl-Meyer Sensorimotor Assessment *Fugl-Meyer et al, 1975*	Limb movement Balance Sensation ROM	based on synergies	3 point ordinal 113 items	30-50 minutes	**	*	—	*	**	**	*
Motor Assessment for Stroke (MAS) *Carr et al, 1985*	Motor performance Tone	functional, task-oriented; motor control theory	7 point ordinal 9 items	15-20 minutes	**	*	—	—	*	—	—
Chedoke-McMaster Stroke Assessment *Gowland et al, 1992*	Impairment: shoulder pain, postural control, limb movement Disability: mobility & walking	based on WHO's ICIDH; & on synergies	7 point ordinal 20 items; (6 impairment; 14 disability)	30-50 minutes	*	*	—	*	*	* (disability inventory only)	*

Key: —: not assessed, * assessed , ** assessed in more than one study & acceptable levels established

(cont...)

TABLE 2 . 1 (cont)

Overview of Characteristics of Stroke Motor Assessments

Scale	Content Domains	Conceptual Framework	Scoring	Administration Time	PSYCHOMETRIC PROPERTIES: Reliability: Inter-rater	Reliability: Intra-rater	Reliability: Internal Consistency	Validity: Content	Validity: Criterion	Validity: Construct	Responsiveness to Clinical Change
Evaluation of the Hemiplegic patient; Bobath approach *Guarrna et al,* 1988	Active movement, Tone, Reflexes, Postural reactions, Sensorium, Pain	based on Bobath's stages of recovery	4 point ordinal 6 components for each limb	30-50 minutes	*	*	—	—	* (u/e subscale only)	*	—
Lavigne Motor Recovery Assessment *Lavigne, 1975*	Movement of u/e, l/e, hand & face Balance, gait & sensation	based on synergies	5 point ordinal for movement (95 items); balance & sensation (38 items) dichotomous; gait (39 items) 3 point ordinal	30-50 minutes	*	—	*	—	—	*	—
Rivermead Assessmement of Motor Function *Lincoln & Leadbitter 1979*	Movement of u/e, l/e, & trunk; Mobility	not specified	2 point ordinal (Guttman) 38 items	15-30 minutes	*	*	—	—	—	—	—
Physical Assessment for Stroke Patients *Ashburn et al, 1982*	Limb movement Mobility	not specified (synergies implied in graded movements)	3 point ordinal for limb movements (27 items) 4 point ordinal for mobility (18 items)	30-50 minutes	*	—	—	—	—	—	—

Key: —: not assessed, * assessed , ** assessed in more than one study & acceptable levels established

2.3 Scale Development

2.3.1 Basic principles of scale development

Measurement can be defined as the process of linking abstract concepts to empirical values (Zeller and Carmines, 1980). The first step, and possibly the greatest challenge in developing an instrument for measuring an abstract concept, is to establish a conceptual framework. This conceptual framework then determines the scope of the measure, the domains included and the method by which they are assessed, as well as the uses and populations for which the measure is appropriate (Green and Lewis, 1986; Ware, 1987). Operationally defining the construct presents the next challenge. For this, scale developers need to clearly delineate the domains to be measured, select the most appropriate items to reflect the construct of interest, and describe in detail how these items will be scored (Michels, 1983; Ware, 1987; Zeller and Carmines, 1980). Having decided what is to be measured, and how to measure it, the next step is to assess how well the theoretical concept is being reflected. This involves establishing the measurement properties. The procedures used for developing and evaluating an instrument, such as a clinical index of motor function, are described below.

2.3.2 Desirable attributes of clinical scales

Several authors have described the various criteria which must be met for a clinical scale to be considered a satisfactory measurement tool (Bergner, 1987; DeVellis, 1991; Jette, 1984; Kane and Kane, 1981; LaRocca, 1989; Law, 1987; McDowell and Newell, 1987; Rothstein, 1985; Streiner and Norman, 1991; Task Force on Standards for Tests and Measurements in Physical Therapy, 1991; Wade, 1992). On the practical side, there are a number of considerations if a scale is intended for use in the clinical setting: 1) ease of administration, recording and scoring; minimal equipment, special training or expertise required, 2) mode of administration, such as self-report questionnaire, interview, or direct observation of performance, 3) respondent burden in terms of time required to complete the test, inconvenience or discomfort, and 4) safety and appropriateness for sample targeted. In general terms, a scale should be quantitative, objective, and comprehensive, yet concise and attractive for use clinically.

In addition to the above practical considerations, an instrument must demonstrate acceptable measurement properties. Reliability and validity are the basic measurement properties typically identified as necessary attributes of an instrument. Responsiveness of the scale for identifying clinical change is another desirable scale characteristic. The

concepts of reliability, validity and responsiveness comprise the psychometric or measurement properties of a scale.

2.3.3 Reliability

Reliability, or reproducibility, refers to whether the measure performs consistently (Feinstein, 1987). It may be examined in relation to the consistency of scoring amongst observers, the stability of the measuring device over time, or the degree of internal consistency of an instrument (Borg and Gall, 1989). The concept of reliability is associated with a lack of random error, where possible sources of variability or random measurement error may include the observer's judgement, the subject's performance, or the discriminatory power of the instrument (Kieth, 1984). The environment and timing of administration can also contribute to variability (Rosenbaum et al, 1990). By the classical test theory, reliability is defined as the proportion of true variance to observed variance (the sum of true and error score variances); by the analysis of variance (ANOVA) model, reliability is the proportion of nonrandom variance, and the proportions of variability due to subjects, raters, and random error can be estimated (Bravo and Potvin, 1991; Zeller and Carmines, 1980). For both models, estimates of reliability vary between 0 and 1, where smaller error variances result in reliability coefficients closer to 1 (Bravo and Potvin, 1991).

A reliable measurement will yield consistent or reproducible results when the characteristic being measured remains stable (Nunnally, 1978). If a scale is found to have high reliability, then changes in scores can be attributed to actual changes in patient status as opposed to human measurement error. If the reliability of scoring is found to be low, it can be enhanced by 1) providing clear instructions or a well defined protocol, 2) basing ratings on clearly stated standardized objective criteria, 3) training raters, 4) using standardized testing conditions, 5) increasing the number of test items, and/or 6) using an average of replicate measurements (Krebs, 1987; Streiner and Norman, 1991).

Several different types of reliability are identified and may differ in importance under different measurement schema. *Inter-rater reliability* refers to the degree of consistency of test scores given by two or more raters. *Intra-rater reliability* refers to the ability of one rater to achieve consistent results with repeated testing under stable conditions. The *test-retest reliability* of a measure refers to the stability of test scores achieved with repeated administration on stable subjects. Thus, while intra-rater and inter-rater reliability are dependent upon the scoring consistency of the evaluators

administering the test, test-retest reliability is dependent upon the characteristics of the test itself (ie. the stability of test performance over time). The duration of the interval between administrations can effect the reliability estimates obtained. This is, therefore, an important consideration in reliability testing; memory effects may confound results if the interval is too short, and the attribute may change if the interval is too long.

For evaluating clinical indices, these forms of reliability are generally assessed by measuring the agreement between observers' ratings of the same performance, or the agreement between two administrations of the test that are separated briefly in time. Use of videotaped performances is one means of ensuring stability of the subject's performance, thus, allowing inferences to be made about the measurement error attributable to the observers. The consistency of raters' ratings should be assessed for any new measurement scenario, as reliability estimates pertain only to the specific conditions under which testing occurs. Also, thorough reports include estimates of the reliability of each individual item, subscales and total scale scores.

Internal consistency, or the homogeneity of a measure, refers to how the test items relate to each other and to the collection of test items as a whole. It concerns the degree to which items intended to measure the same characteristics receive similar ratings when tested. The internal consistency of a measure is an important consideration when an instrument is made up of a number of items which are summed to produce a total score which is intended to measure an abstract construct, such as motor performance. This form of reliability is typically estimated by item to item or item to total correlation, or by Cronbach's coefficient alpha, which represents the average correlation among items within a test (Cronbach, 1951). A high value of alpha (greater than 0.80) suggests that the overall score attained using the multiple items is a better reflection of the underlying factor than can be obtained by individual items alone (Nunnally, 1978).

2.3.4 Validity

Validity refers to the extent to which an instrument measures what it is intended to measure. In other words, it is the fidelity of the inferences made when interpreting a measure. Validity, thus, is determined by the extent to which the underlying phenomenon or concept of interest is reflected by the measure, and by the accuracy of the measure. As with reliability, validity is a matter of degree. Both random and systematic error or bias can contribute to a reduced validity, hence reliability is a prerequisite of validity. Validity is also context specific; though an instrument may be

valid in one setting or for one group of individuals, it may not be valid in other circumstances (Nunnally, 1978). A measure is only validated for a given purpose, and must be reassessed for any novel application of the measure, thus, in actuality it is the application of the instrument which is validated, and not the instrument per se. The process of validation requires accumulating empirical evidence to support that the instrument actually measures what it purports to measure. The conceptual forms of validity discussed in the literature include *content* validity, *criterion* validity (also referred to as *concurrent* and *predictive* validities) and *construct* validity.

Content validity refers to the representativeness, or the degree to which the items reflect the full domain of interest. Thus, content validity relates to whether the domain is adequately covered, and how well the items making up the instrument represent the spectrum of possible items which could be chosen to measure the domain. Measures that tap the full range of a domain will yield a wide and appropriate distribution of scores (Ware, 1987). The basic processes involved in obtaining content validity are 1) specifying the domain of content, and 2) constructing and/or selecting items associated with the domain of content (Zeller and Carmines, 1980). Several steps are advocated to produce a scale which possesses content validity, including the following: 1) a thorough review of the literature related to the content area, 2) clearly defining the objectives and target population, 3) questioning patients and knowledgeable professionals to formulate representative test items and to generate an item pool, 4) preparing and trialing a prototype, 5) repeated reviews and refinements of the instrument by consensus of a panel of expert judges, and 6) statistical analysis to examine whether items group together as hypothesized and to identify how best to construct a composite index (Borg and Gall, 1989; Green and Lewis, 1986; Guyatt et al, 1987; Jolliffe, 1992; Streiner and Norman, 1991). Factors to be considered in selecting items and scales include the goals of the measurement, practical considerations, and the population of interest (Ware, 1987).

The 'face validity' of a measure is based on an expert's opinion regarding the content of the scale, and the 'consensual validity' is based on the opinions of a panel of experts; these are two sub-types of content validity. Two techniques for obtaining consensus that have been widely used in instrument development are the "nominal group technique" and the "Delphi process"(Delbecq et al, 1975). Briefly, the nominal group technique involves a structured meeting that follows a prescribed format. With this technique, relevant ideas are contributed by each participant. These ideas are

recorded and then discussed in an effort to elaborate the meaning and to generate further suggestions that may have been missed. Finally, a vote is taken on each suggestion. The decisions regarding implementing the suggestions are based on *a priori* criteria. The other consensus method, the Delphi process, involves soliciting and collating judgements through a series of mailed questionnaires. The information derived from earlier responses is summarized and included with each questionnaire. The process is repeated until the responses from participants approach consensus (Dalkey, 1967). Studies suggest that the two consensus methods described yield comparable results, however, the results obtained from the nominal group technique are more immediately available (Delbecq et al, 1975). For these and similar consensus methods, an agreement score can be calculated on the extent to which experts agree on the appropriateness of the items and domains of which the instrument is comprised, thus providing an empirical measure of consensual validity (Green and Lewis, 1986). The attributes appraised as part of content validation may include: omission of important variables, inclusion of inappropriate variables, weighting of variables, and clarity or sensibility of presentation (DeVellis, 1991; Feinstein, 1987).

Criterion validity refers to the relationship between the scores obtained using the scale and another external criteria or measure of the same phenomenon. Two sub-types of criterion validity have been described in relation to the timing of the assessments; these are concurrent and predictive criterion validity. Concurrent criterion validity refers to the relationship between the scale score and the criterion measure assessed at the same point in time, and is frequently reported as a correlation between the test scores. Predictive criterion validity is evaluated by the correlation of scale score with a criterion measure or event which occurs at a later date. Evaluation of predictive validity requires a longitudinal study, or retrospective evaluation of measurements obtained at different points in time.

Ideally criterion validity is assessed against a "gold standard" or universally accepted measure of the phenomenon of interest, if such a measure exists. If the measure being developed is found to correlate highly with the accepted measure, yet is less invasive, less expensive, and/or easier to administer than the so-called "gold-standard', it may then be utilized in place of the standard measure. Unfortunately, many concepts are abstract and have no generally accepted "gold standard" to serve as criterion measure, thus, careful selection of a suitable related measure which approximates the phenomenon of interest offers the next best alternative. The strength of the association

found between the measures will be affected by the adequacy or representativeness of the criterion measure selected; caution must therefore be exercised in interpreting the results of criterion validity studies.

Construct validity relates to testing assumptions that an instrument measures a specific concept, and that the given concept performs according to theoretical expectations (Zeller and Carmines, 1980). Measures of abstract concepts that are not defined adequately by a criterion measure, or a universe of content, must possess construct validity. Nunnally (1978, p84) has eloquently expressed the difficulties encountered in validating a measure of an abstract phenomenon, saying that "the degree to which it is necessary and difficult to validate measures... is proportional to the degree to which the variable is concrete or abstract." For constructs such as motor function, the task of construct validation is an exercise in creativity and logic; the more ways that construct validity of a new measure is tested, the greater the confidence one can have in the performance of the measure (Del Greco et al, 1987).

In construct validation, assumptions are made about the dimensions which a composite measure, consisting of multiple items, is intended to reflect. Factor analysis is a means of empirically testing the strength of relationships between items or domains of an instrument; items intended to measure the same underlying concept correlate highly ("converge and load high") on one factor (Cronbach, 1951). Assumptions are also made about a network of interrelationships amongst constructs external to the instrument. Hypotheses may be formed about the behavior of test scores in numerous situations, and correlations between associated measures examined for predicted direction or magnitude of relationship. The correlations used in criterion-related validity assessment can also be used for construct validation, which may result in some confusion between these two issues. The investigators intent, the theoretical explanation of a construct versus comparability to related measures, is where the difference lies between these two forms of validity testing (DeVellis, 1991).

The process of scale validation is a cumulative task; it is only with accrual of evidence from repeated and varied validity assessments that a clinical scale gains credibility. Furthermore, validity of the measurement tool is not sufficient to obtain valid results in a study; validity of design and analysis are also crucial (Nunnally, 1978).

2.3.5 Responsiveness

Responsiveness, sometimes referred to as sensitivity to change, is the ability of a scale to detect clinically important changes in the characteristic of interest over time. In studies of treatment efficacy, where recovery is expected to occur in small but clinically important increments, a scale which includes items that the treatment program is expected to impact upon will likely be the most responsive to change in patient status (Bergner, 1987). By including items that effectively detect changes, and by offering a range of response options, the measures potential to identify finer gradations of change is enhanced (Rosenbaum et al, 1990). Thus, in order to be highly responsive, a scale must be reliable, and must be adequately scaled for the desired degree of discrimination. As patients may experience improvement in some areas, while in other areas no change or a deterioration may occur, it may be of interest to examine change scores item by item to reveal individual items reflecting the greatest changes.

A number of empirical methods have been suggested to measure a scale's responsiveness (Deyo and Centor, 1986; Deyo and Inui 1984; Liang et al, 1985; MacKenzie et al, 1986; Sackett et al, 1977). One means is to simply examine the range of changes in scores in terms of expected direction and magnitude of change, taking into consideration the variance present in the measurements. Correlations between score changes measured by a number of related instruments (ie. admission to discharge change scores) is another means used to estimate the relative responsiveness of an instrument. It may be possible to examine the within person change following an intervention of known efficacy; within-patient score changes (ie. before and after the intervention) for various measures can be linearly transformed to a common range of possible values (eg. out of 100) or z scores can be calculated to compare measures. Liang et al (1985) used t-statistics to reflect scales' relative responsiveness as compared to another measure's t-statistic. The ratio between t-statistics is essentially equal to the ratio between coefficients of variation, except that the t-statistic includes a sample-size factor that corrects for unequally sized samples in cases where data is missing. Finally, a model has been proposed whereby a clinical scale can be looked at as a diagnostic test, and sensitivity and specificity calculated from a two by two table where high and low scoring groups of subjects (with pre-specified score cut-off points) are compared with an external criterion based on clinical judgement of patient status.

An issue related to the sensitivity of the scale to clinically important change, and one that impacts on sample size requirements when the scale is used for clinical studies, is

the efficiency of a measure. Efficiency may be estimated by effect size or coefficient of variation (CV). Effect size is calculated as the ratio of the difference between group mean change scores for a variable (the "before and after" measures), divided by the sample standard deviation (SD) of the variable at baseline; the coefficient of variation is defined as the ratio between group SD divided by the respective group mean. Effect sizes range from 0 to 1, with larger values representing greater efficiency in terms of statistical power for a given sample size, or smaller sample size requirements for a given power (Kazis et al, 1989; Ottenbacker, 1989). Conversely, the lower the CV, the more efficient is the measure, as this indicates that the SD, or variability, is small relative to the mean. The relative efficiency of a measure (in terms of its statistical properties) is one additional aspect that should be determined and reported in the process of scale development, as this information is of use to those attempting to choose the optimal instrument for use in a study.

2.4 Directions From the Literature

In summary, the evaluation of the efficacy of rehabilitation programs and the development of measurement tools to meet this aim are research priorities. While motor recovery following stroke is an important focus of physiotherapy assessment and treatment, a consistent and comprehensive operational definition of motor function has yet to be achieved. None of the currently used clinical measures of motor function has widespread utility. Careful instrument construction, and detailed evaluation of the measurement properties outlined in this review, are important and necessary steps in the process of scale development.

CHAPTER 3

OBJECTIVES AND METHODS

3.0 Objectives

This thesis project has two global objectives. They are to improve and validate the content of the STREAM, and to establish preliminary estimates of the reliability of this clinical instrument.

To this end, the specific objectives are:

1) to revise the STREAM so that the items are appropriate, representative and comprehensive for the evaluation of motor recovery following stroke (**content validity**);

2) to determine the extent to which pairs of therapists concur on the scoring of the items of the revised STREAM and on the total score (**inter-rater reliability**);

3) to examine the extent to which the items included in the revised STREAM relate to each other, and to the group of items as a whole (**internal consistency**) and;

4) to assess the reliability of scoring of videotaped performances of STREAM across occasions and across raters (**intra- and inter-rater reliability**).

3.0.1 Overview of study design

To meet the specific objectives, three separate sub-studies were conducted. The methods relating to each of these sub-studies are described below under the sub-headings: 1) 'developing and refining the content of STREAM', 2) 'the direct observation reliability study', and 3) the 'videotaped assessments reliability study'. Objective one was achieved in two phases. First, two consensus panels were convened to make recommendations to improve the content validity of the original version of STREAM, and to prepare a test version of the instrument. Next, this intermediate test version of STREAM underwent initial evaluations of reliability and internal consistency to identify items that should be eliminated, and item reduction was done to produce the completed STREAM. For the second objective, a reliability study involving pairs of raters, who directly observed and assessed patients, was carried out at the Jewish

Rehabilitation Hospital (JRH). To realize the third objective, the associations between the items were examined using data accrued in the direct observation study, as well as additional data collected on patients in the lower end of the scoring range. And, to accomplish objective four, a reliability study, using videotaped assessments viewed and rated on two occasions, was done. The procedures for each sub-study are detailed separately. However, because there were many similarities in the statistical methods used for each of the sub-studies, the methods for analysis are presented together in section 3.4 to facilitate comparisons and avoid redundancy.

3.1 Developing and Refining the Content of STREAM

The focus of this sub-study was content validation (objective one). This was accomplished in two stages: 1) the items and scoring of the original STREAM were revised using consensus methods, and 2) the number of items were reduced, based on a preliminary evaluation of inter-rater agreement and internal consistency of the test version of STREAM, and on the opinions expressed by the consensus panels.

3.1.1 Selecting the consensus panel participants

The participants on the consensus panels were therapists working in a variety of clinical settings, dealing with patients in all phases of stroke rehabilitation. The first consensus panel was recruited from ten Montreal area health care facilities, and included therapists who responded to a notice that was sent to the physiotherapy departments. Eleven physical therapists with more than one year of experience working with stroke patients participated. The second panel, involving nine therapists, was assembled from therapists named by members of the first panel. As the therapists at the JRH had experience using the original STREAM clinically, a representative from the JRH physiotherapy staff was included on each panel.

3.1.2 Developing the test version of STREAM

Prior to the meeting of the first panel, each participant was mailed a copy of the original version of STREAM and a letter indicating their role in revising the STREAM (Appendix 3.1.1). The panelists were instructed to review the STREAM, and to consider changes related to the items included, the wording, and the scoring scheme. Each participant was provided with a form on which to list their recommended changes (Appendix 3.1.2). The panel members were asked to familiarize themselves with the STREAM prior to attending the meeting, by using it to assess two patients. At the meeting, a brief introduction to the conceptual framework for STREAM and an

overview of the study was given. A videotaped assessment of a patient was also shown.

The procedures for orienting the second panel were identical to the first, except that participants were asked to review the first revision of STREAM (Appendix 3.2.2). In addition, prior to attending the panel meeting, they were asked to consider three different scoring schemes (Appendix 3.2.4), or recommend a fourth. And, there was a slight modification of the opening remarks of the meeting in that this second panel was given the results of the content verification survey and the outcome of the work of the first panel.

The procedures followed during the two consensus panel meetings, of approximately three hours duration, closely approximated the '*nominal group process*' described by Delbecq et al (1975). Briefly, the process involved the following steps: 1) individual generation of suggested changes, 2) compiling these suggested changes, 3) serial discussion of each suggestion, and 4) voting on the importance of each suggestion. Therapists presented items from their prepared lists of suggested changes. Each suggestion was recorded on a large chart. Three separate charts were made: one for suggested additions and deletions, one for clarifications or modifications to existing items, and one for changes in scoring. The process continued until no new suggestions were raised. Next, each of the suggested changes was discussed in turn. The goals of this discussion were to elaborate on or to clarify the meaning of the suggested changes, to air experiences and strong opinions regarding the suggested changes, and to add new items that emerged through the discussion. Finally, the panel members were asked to independently rate each of the suggested additions and deletions in terms of importance, using the same five point ordinal scale that was used during the preliminary content verification survey (1:crucial; 2 :very important; 3:moderately important; 4:neither important nor unimportant; 5:unimportant). The rating form is provided in Appendix 3.2.5. For each of the suggested additions or deletions, the number of therapists (out of eleven for the first panel; out of nine for the second panel) giving a particular rating was tabulated. An *a priori* criterion of greater than 50% of the therapists rating an item addition or deletion as important (ratings of 1, 2, or 3) was set for implementing the change. In addition, the therapists rated items regarding clarifications or scoring changes simply as agree or disagree, and a majority vote was required for the suggested change to be implemented.

3.1.3 Follow-up questionnaire on the revised scoring scheme

As consensus had been reached regarding the items to include, but had not been fully reached regarding the scoring of STREAM, the comments from the two panels and from the content verification survey were collated, and a new scoring scheme was devised based on these comments. A brief questionnaire (Appendix 3.2.6) was sent to the participants of both panels, in order to confirm the appropriateness of this revised scoring scheme.

3.1.4 Criteria for item reduction

Once consensus was reached on the content and scoring of STREAM, the test version of STREAM was ready for further refinement. The focus of this stage of STREAM's evolution was to identify items that should be eliminated. The criteria for item reduction were based on a combination of statistical information, clinical relevance, and face validity. The statistical aspects that were to influence the selection of items included the relationship of each item to the collection of items as a whole (internal consistency) and the reliability of scoring of the individual items. Data from the direct observation reliability study, including inter-item, and item to subscale and total correlations, and item kappa statistics for the test version of STREAM, were used in the process of item reduction. The methods used for subject and rater selection, training the raters, and assessing the patients in the direct observation study are described in section 3.2; the statistical methods are presented in detail in section 3.4. The specific criteria for item reduction are presented in Table 3.1, and the justifications for these criteria follows. Items would be eliminated if they exhibited two or more of the characteristics summarized in this table.

TABLE 3.1
Criteria for Item Reduction

1) 1/3 (14/42) or more of inter-item correlations outside the range of 0.4--0.85

2) correlation(s) of more than 0.9 with item(s) measuring the same limb segment or a closely related mobility task

3) an item-to-total correlation outside the range of 0.4--0.85

4) an item Kappa statistic below 0.6

5) not considered important for measuring motor recovery (not recommended by consensus panels) or limited by factors other than motor recovery (eg. ROM, pain)

Thus, items with item-to-item and item-to-total correlations in the mid-range, thereby reflecting homogeneity without redundancy, would be retained, while items that were not associated with the test as a whole would be eliminated. Similarly, items found to be highly correlated with other items measuring related movement components (eg. shoulder abduction and reaching to the top of the head, shoulder flexion to 90 degrees and full shoulder flexion, taking steps or turning to either side, or climbing and descending stairs) would be eliminated to avoid redundancy. Items that could not be reliably scored would be dropped. Also, any items that appeared to be highly influenced by constructs other than motor recovery, or were not related to the subscale (as reflected by low correlations between item and subscale score), would be questioned on the grounds of dubious face validity or clinical relevance. However, in order to maintain the face validity of STREAM, an item that did exhibit two of the criteria would be retained if it measured a unique component of motor recovery that was not incorporated by any other item, and it had been considered important by the consensus panel members.

3.2 Methods for the Direct Observation Reliability Study

The data obtained in this sub-study contributed to attaining the first three objectives. An evaluation of inter-rater agreement on the scoring of items and internal consistency of the test version of STREAM was used to complete the content validation process as described above (objective one). Further appraisal of the data, including only the items retained, provided estimates of inter-rater agreement (objective two) and internal consistency (objective three) for the completed instrument.

3.2.1 Selecting the subjects

The study was carried out at the Jewish Rehabilitation Hospital (JRH), a 120 bed rehabilitation hospital with a forty bed stroke unit that admits approximately 200 stroke patients per year. A convenience sample of twenty cooperative persons admitted to the JRH for active physical rehabilitation for the treatment of motor dysfunction following stroke were chosen. These people were selected to represent a wide range of motor dysfunction in terms of degree of disability, age, and time since stroke. We excluded patients with any major comorbid conditions which interfered with motor function, or its assessment, such as: a neurological condition in addition to the stroke; a severe comprehension disorder; marked bilateral motor or sensory impairment; amputation of a limb; or severe rheumatoid arthritis.

In addition, for analyzing the internal consistency of both the test and the completed versions of STREAM, scores for a total of twenty-six individuals were used. This sample included the twenty subjects identified above, and an additional six patients whose scores were expected to be in the lower end of the range. The six additional low level STREAM scores were collected because none of the twenty rehabilitation patients who participated in the JRH reliability study had received scores on the test version of STREAM below thirty. Therapists that were participating in the videotaped assessments reliability study, and who were, therefore, familiar with administering the STREAM, volunteered to provide us with scores on STREAM for their low level patients. These low level STREAM scores were collected at four different facilities, including: one long term care (LTC) facility (Cote-des-Neiges Hospital), and three acute care hospitals (Montreal Neurological Institute, Lakeshore General, and Reddy Memorial).

3.2.2 Ethical considerations

Ethical approval was obtained from the ethics committees at both the JRH and the School of Physical and Occupational Therapy, McGill University. The following conditions were imposed by the JRH ethics committee: 1) patients were to be assessed after regular therapy hours, so that there was no interruption of therapy time, 2) raters were not to give feedback to patients regarding their performance, but were to direct patients' inquiries to their treating therapists, and 3) only patients who were not already involved in another research study were to be recruited. Also, informed consent (Appendix 5.1) was to be obtained prior to a patient's participation in the reliability study.

3.2.3 The raters

The raters for this study were four therapists recruited from the physical therapy department at the JRH, one experienced therapist from the Montreal General Hospital, an acute care institution, and the author (KD). All raters had at least one year of experience working with stroke patients as a physiotherapist.

3.2.4 Training the raters

One week prior to commencing the reliability study, therapists participated in a two hour training on the test version of STREAM. This training session included a discussion of revisions made to STREAM, and a presentation of a videotaped patient assessment using the test version of STREAM. Each therapist was given a copy of the

STREAM and scoring instructions to review, and was asked to practice administering the test to at least two patients prior to the study.

3.2.5 Assessing the patients

Consenting patients were evaluated on the test version of STREAM simultaneously but independently by a pair of raters. One therapist performed the assessment and the other therapist observed. Each therapist rated between four and eleven patients (mean 6.7 patients); in approximately half of the rating sessions they performed the assessments, and were the observer in the other cases. The process of assigning raters to assess subjects was based solely on the availability of a given rater on any given day, with the aim of having each therapist evaluate at least four patients over the course of the study. During the rating sessions the raters were unaware of each others scores, and they did not discuss the patient's performance. STREAM scores were collected on the scoring forms provided (Appendix 6.1). The rating sessions for the twenty subjects were carried out over a three month period, and required from fifteen to forty-five minutes each.

3.3 Methods for the Videotaped Assessments Reliability Study

The data obtained in this sub-study contributed to attaining the fourth objective. That is, estimates were obtained for the intra- and inter-rater agreement on the scoring of videotaped performances of STREAM.

3.3.1 The raters

The twenty raters for this study were recruited from Montreal area health care facilities by having the consensus panel participants inform their colleagues about the upcoming reliability study, and by sending notices to hospital physical therapy departments explaining the study. The raters were selected to cover a wide range of levels of expertise and experience, with a minimum of six months of experience working with stroke patients. An additional requirement for participating was that they were not familiar with any of the four videotaped subjects.

3.3.2 Videotaping the STREAM assessments

Four patients, who had participated in the direct observation study, were selected to participate in this phase of the study. These patients were reassessed on STREAM, and the assessments were videotaped. The same therapist performed the assessments

for each of the videotapes. These four videotaped performances clearly showed each test item being performed, and displayed a range of motor function.

3.3.3 Ethical considerations

Patients who agreed to be videotaped, signed a release form (Appendix 5.2) indicating their agreement to have their assessment shown to others for the purposes of teaching or research. Additional burden to subjects was minimized by utilizing these videotaped performances for this portion of the reliability study, as repeated assessments by twenty raters would have been excessive. Confidentiality of the videotaped subjects' names was maintained.

3.3.4 Rating the tapes

Two weeks prior to the viewing session, the raters were given a copy of the test version of STREAM and the scoring instructions to study, and they were asked to use the STREAM for assessing at least two patients. At the beginning of the viewing sessions, written instructions for the procedures to be followed during the viewing (Appendix 6.3) were given to the raters and discussed. For practice, a sample videotaped performance (not used in the reliability study) was shown to the raters.

The twenty raters, divided into two groups of ten for convenience of viewing, simultaneously and independently evaluated the four videotaped assessments. During the rating sessions, therapists recorded their ratings on the forms devised for the study (Appendix 6.1), and no discussion of the scoring of any of the items was permitted. Items were allowed to be replayed up to three times upon request (although this rarely occurred), as some of the smaller movements were more difficult to see on video, and because, when the testing is done in the clinical setting, the patient is permitted up to three attempts to perform a test item. The raters viewed each of the four videotaped performances on one occasion, and on a second occasion approximately one month later. The videotapes were presented in a random order at each session.

3.4 Analyzing the Data from the Reliability Studies

3.4.1 Overview of the data analysis

In this section, the procedures used for data collection, and the software used for processing the data are described first. Next, the statistical methods used in the initial evaluation of the test version of STREAM are presented. These include an analysis of internal consistency, and agreement on the scoring of the items. Finally, the

characteristics of the statistical tests used in evaluating the reliability of the completed STREAM are presented in detail. It should be noted that all of the items included in the test version of STREAM were scored during the reliability studies, however, only the kappa statistics reflecting rater agreement on item scoring, and the internal consistency, were evaluated for the test version. Following the item reduction, a more indepth evaluation of the rater agreement for the scoring of the remaining items, and a second analysis of internal consistency, were carried out for the completed version of STREAM. The analyses of rater agreement were parallel for the two reliability studies, except that only inter-rater agreement was evaluated in the direct observation study, while estimates of intra- and inter-rater agreement were derived from the videotaped assessments study.

The STREAM scores were collected on forms devised for the reliability studies (Appendix 6.1), entered on a spread sheet, and inspected for accuracy. SAS statistical software (SAS Institute, 1985) was used to compute Pearson's correlations, Cronbach's alphas, cell frequencies and Kappa statistics, and signed rank statistics. The GENOVA version 2.1 program (Crick and Brennan, 1983) was used to obtain GCCs for subscale and total scores.

3.4.2 Internal consistency of the test version and completed STREAM

An assessment of internal consistency is concerned with the homogeneity of an instrument. That is, internal consistency reflects the degree to which items included relate to each other and to the instrument as a whole. Although internal consistency is traditionally considered internal reliability, it also provides information for content development. For example, the inter-item, item to subscale and item to total correlations, and alpha coefficients, provide indications as to whether a given item should be retained or removed from the instrument.

For evaluating the internal consistency of STREAM, our sample included twenty-six individuals (twenty from the JRH, and six additional low scoring patients) covering the full range of possible scores, so that the whole population for which STREAM is intended was represented. We used the twenty scores on STREAM given by the assessing therapists in the direct observation reliability study, rather than the scores given by the observing raters, because the assessors' scores more accurately reflect the typical clinical situation where scores are obtained by directly assessing and rating a patient, and not by observing the assessment.

Three parameters relating to internal consistency were calculated for both the test and completed versions of STREAM. These included: 1) the Pearson's correlation coefficients for each possible pair of items included in the revised STREAM (inter-item correlations); 2) the correlations between the scores for individual items and the subscale and total scores, calculated by omitting that item (item to total correlations); and 3) Chronbach's alphas (Chronbach, 1951) for each subscale, and for the STREAM as a whole (average inter-item correlations).

No specific guidelines for interpreting inter-item and item to total correlations are given in the literature on internal consistency and item reduction, as the magnitude of these correlations varies with the total number of items included in the instrument. As a generalization, however, inter-item correlations in the moderate range (eg. 0.4--0.85) are considered optimal, as within this range items are probably measuring related aspects of a trait, but are not so highly related as to be redundant (DeVellis, 1991; Feinstein, 1987; Streiner and Norman, 1991). Similarly, low item-to-total correlations (eg. below 0.4) indicate that an item is measuring a different trait than the rest of the instrument as a whole; very high item-to-total correlations (eg. above 0.85) suggest that an item is contributing information about the trait that has already been provided by the other items. Obviously, items within a subscale are expected to be measuring the same trait, so item to subscale total correlations should be high. A Cronbach's alpha is generally considered acceptable if greater than 0.7, good if greater than 0.8, and excellent if above 0.9. However, since Cronbach's alpha is influenced by the total number of items included in an instrument, and increases in value if related items are added, alpha must be interpreted accordingly. As a general guideline, for an instrument to be clinically useful for measuring a particular trait, alphas of at least 0.9 are recommended (Feinstein, 1987).

3.4.3 Inter-rater agreement for items on the test version of STREAM

The agreement between raters for scoring items on the test version of STREAM in the direct observation study was described using the index of crude agreement (the total percentage of subjects in which paired scores agree precisely), expected agreement, and Cohen's Kappa statistic (Cohen, 1960). The **Kappa statistic** is defined as the ratio between observed and expected agreements, and the difference between maximum possible agreement and the expected agreement. It is interpreted as the proportion of agreement between two judges rating n subjects after chance agreement has been removed (Cohen, 1960). The Kappa statistic is prevalence dependent; it is influenced

by the distribution of scores, the variability among subjects, and the number of rating categories (Feinstein and Cicchetti, 1990; Soeken and Prescot, 1986). Therefore, if Kappa is to be used to calculate agreement, it is essential that subjects with a wide range of capabilities are included. Kappa ranges from zero to unity; if the observed agreement is equal to that expected by chance, kappa is zero. Values of Kappa greater than 0.75 are considered to represent excellent agreement beyond chance, while Kappas of 0.4 to 0.75 represent moderate reliability (Landis & Koch, 1975).

For measuring the agreement between items scored on an ordinal scale, weights can be assigned that correspond to the degree of importance of the observed disagreements. For example, linear weights are typically assigned to each category for scales in which disagreements between the different categories are all deemed to be of the same magnitude. For some measures, however, the degree of importance of disagreements between categories may be viewed as varying exponentially. The observed and expected agreements obtained when appropriate weights are attributed to the disagreements between categories can then be used to calculate a **weighted Kappa** statistic (Cohen, 1968; Kramer and Feinstein, 1981). The linearly weighted Kappa reflects the chance corrected agreement where the disagreements between scoring categories are considered to be additive, or on a continuum; the exponentially weighted Kappa reflects a chance corrected agreement where the disagreements between categories are compounded, or magnified by the distance between one another. When Kappa is determined by using exponential weighting, quadratic weights are recommended, and the quadratically weighted Kappa values obtained are conveniently equivalent to the ICCs for the same data (Fleiss and Cohen, 1973; Streiner and Norman, 1991). Thus, quadratically weighted Kappa statistics were calculated for the agreement on the item scores (scored on an ordinal scale of 0 to 2 for limb movements, and 0 to 3 for basic mobility items), as it was felt that differences in scores that were two categories apart were markedly more serious than differences in scores that differed by only one category.

3.4.4 Inter- and intra-rater agreement for items on the completed STREAM

As described above, kappa statistics were used to reflect agreement on the scoring of items in the direct observation reliability study. However, as Kappa statistics could not be derived for the individual item scores of the videotaped assessments reliability study, due to insufficient variability in scores with only four subjects involved, non-parametric signed rank statistics were computed to evaluate the agreement on item scores, and to

identify trends in the scoring of the items. Signed rank statistics were also computed for the direct observation reliability study, in order to allow a comparison between the two reliability studies. The **Wilcoxon matched-pairs signed rank statistic** (Wilcoxon, 1945) is the non-parametric equivalent of the paired-sample t-test. It examines the differences between pairs of scores, which under the null hypothesis are assumed to have a median difference of 0 (Colton, 1974). If the signed rank statistic is significant, this indicates poor agreement due to a tendency to score either consistently higher or lower on the different rating occasions, or by either the assessing or observing raters. The SAS univariate procedure produces a centered signed rank statistic, which is computed using the sum of the positive ranks, and which used average ranks for tied values (SAS Institute, 1985).

In addition, as was done for describing rater agreement on the scoring of items in the preliminary analysis, the rater agreement for the scoring within category **'1'** (ie. between the sub-categories 1a, 1b, and 1c) for the direct observation study was described using the index of crude agreement, expected agreement, and Cohen's Kappa statistic. The scoring within category **'1'** was not considered to be ordinal, but rather to be qualitative and nominal. Therefore, unweighted Kappa statistics, which treat all disagreements equally, were calculated for describing the agreement within scoring category **'1'** for the direct observation study. Because meaningful kappa statistics could not be computed for the videotaped assessments study due to insufficient variability, only the distribution of agreement for scoring within category **'1'** was presented for this study.

3.4.5 Rater agreement for subscale and total scores on the completed STREAM

Scores for the items that were retained in the completed instrument were summed to produce subscale and total scores on STREAM and this data set was used for the following analyses. Because there were some items scored **'X'** in the direct observation reliability study, the subscale and total scores were transformed to scores out of 100 (see formula presented in Appendix 8.3). However, there was no missing data in the videotaped assessments study; the limb subscales are thus scored out of a maximum of twenty each, mobility is scored o t of a maximum of thirty points, and maximum total STREAM score is seventy.

For describing the agreement on subscale and composite scores, **Generalizability Correlation Coefficients** (GCCs) were calculated (Cronbach et al, 1972). For the direct observation study, these statistics reflected agreement between raters, while for

the videotaped assessments study GCCs described agreement both within and between raters. GCCs, based on the generalizability theory, are analogous to traditional reliability coefficients, except that GCCs reflect not only the magnitude of the error as would a traditional reliability coefficient, but also attribute the error to a specific source (Crick and Brennan, 1983; DeVellis, 1991; Streiner and Norman, 1991); they resemble the ICC, in that both GCCs and ICCs range from 0 to 1, and are based on an ANOVA model. The closer the GCC is to unity, the greater the generalizability or reliability. Reliability coefficients above 0.80 are generally considered acceptable for tests that are used to make decisions about a group or for research purposes, however, for tests used to make judgements about individuals, coefficients of greater than 0.95 have been recommended (Helmstadter, 1964; Nunally, 1978; Weiner and Stuart, 1984).

In both the direct observation reliability study and the videotaped assessments study, various subjects and raters (called 'facets' in the parlance of the generalizability theory) contribute to the variability of the error terms. In the videotaped study, occasions are additional facets contributing to the variance in scores. For each study, several different GCCs were calculated, each focussing on different facets or 'objects of measurement', including subjects, raters, and occasions. The 'object of measurement' in a generalizability study is the facet that the researcher wishes to make generalizations over-- ie. the specific aspect of reliability under scrutiny. In addition, GCCs were calculated for each variable under both fixed or random conditions. To examine inter-rater reliability, where determining the generalizability across raters is the goal, the object of measurement is raters, and the variance in scores that is contributed by occasion is excluded from the sources of variance used to calculate the GCC. To examine intra-rater reliability, where the goal is to be able to generalize across occasions, the variance in scores that is contributed by raters is omitted from the equation. Thus, by examining the effects of subjects, raters, (and, for the videotape study, occasions as well) simultaneously, we determined the extent to which (and under what conditions-- ie. fixed or random) we can generalize across each facet.

In addition to GCCs, signed rank statistics were computed on subscale and total STREAM scores for the two reliability studies. As with the analyses of the agreement on the individual items, this non-parametric statistic served to reflect the magnitude and significance of observed disagreements, and to identify trends in the scoring by the assessors and observers in the direct observation study, and between the two viewing sessions in the videotaped assessments study.

3.5 Summary of Methods

This study consisted of two distinct and sequential phases: content validation, and preliminary testing of reliability. The content verification survey, described previously in section 1.2, was a preliminary step in the process of content validation. Subsequently, two consensus panels, involving a total of twenty physical therapists working in a wide range of settings, made recommendations to improve the content validity of STREAM. Next, preliminary evaluations of the internal consistency and reliability of scoring the items of the test version of STREAM were carried out to identify items that should be dropped from STREAM. Item reduction, based on the criteria outlined in Table 3.1, produced a more streamlined STREAM, which subsequently underwent more detailed evaluations of reliability. Two separate reliability studies were conducted: 1) a direct observation reliability study, performed in the clinical setting, involving twenty stroke patients, and pairs of physical therapists from a pool of six raters, and 2) a study using four videotaped assessments, involving twenty physical therapists as raters, and two viewing and rating sessions. Preliminary evaluations of the data from the direct observation reliability study provided the statistical information required for item reduction. The internal consistency of the STREAM (test version and completed) was evaluated using scores on STREAM for twenty-six individuals, including the scores given by the assessors in the direct observation reliability study, and from six additional stroke patients with scores on STREAM in the low end of the range. The reliability of the completed STREAM was analyzed using only the portion of data pertaining to the items retained following item reduction. The data from the direct observation reliability study was used to estimate inter-rater reliability for the STREAM and its subscales. The data from the videotaped assessments reliability study provided estimates of both intra- and inter-rater reliability.

CHAPTER 4

RESULTS

4.0 Introduction

The results section is presented in a sequence that is parallel to the organization of the previous chapter on methods. That is, the details regarding the development of the content of the test version of STREAM are given first. These are followed by the results of the analyses done on the test version, including the internal consistency and Kappa statistics of the original forty-three items. Next, the process of item reduction, and the completed thirty item instrument, are presented. The remainder of the results relate to the thirty item STREAM, including the internal consistency of this completed version of STREAM, the inter-rater agreement from the direct observation study, and the intra- and inter-rater agreement from the videotaped assessments reliability study.

4.1 The Test Version of STREAM

4.1.1 Characteristics of the panel participants

Panel one consisted of two therapists working in acute care settings, seven from in-patient rehabilitation facilities, and two from out-patient rehabilitation settings. These therapists had from eight to thirty-two years (mean 15.2 years$\pm$8.4) of experience as physiotherapists, and from three to twenty years (mean 8.8 years$\pm$6.0) of stroke related experience. Panel two included three therapists working in acute care settings, three in in-patient rehabilitation facilities, one in an out-patient rehabilitation setting, and two therapists from long-term care facilities. The panel two therapists had from two to twenty-four years (mean 9.6 years$\pm$6.7) of experience as physiotherapists, and from two to seventeen years (mean 6.6 years$\pm$5.0) of stroke related experience. Appendix 3.3 outlines the characteristics of the panel participants in detail.

4.1.2 The evolution of STREAM items

The changes that were suggested by each of the consensus panels, and the rationale for adding, deleting, or changing the items are briefly summarized in Appendix 4. The evolution of the items included in the STREAM, and the stage of refinement (including: the original version, following panel one, following panel two, test version of STREAM used in the reliability studies, and the completed thirty item instrument) during which each item was added or deleted is shown in Table 4.1. The minor modifications made to several items are also noted (bracketed). Nine items which had

TABLE 4.1

The Evolution of STREAM Items

Item Description	Original STREAM	Revision 1 First Panel	Revision 2 Second Panel	Revision 3 Test Version	30 Item STREAM
LIMB MOVEMENTS					
Upper Extremity					
Protracts scapula in supine				*	+
Extends elbow in supine				*	+
Shrugs shoulders		+	+	+	+
Affected hand touches ...	+ (ear)	+ (forehead)	+ (top of head)	+ (top of head)	+
Affected hand touches sacrum	+	+	+	+	+
Raises affected arm to forward-horizontal	+	+	+	+	--
Raises affected arm to lateral-horizontal	+	+	+	+	--
Raises affected arm to full flexion	+	+	+	+	+
Supinates and pronates forearm	+	+	+	+	+
Actively closes the affected hand	+	+	+	+	+
Actively opens the affected hand	+	+	+	+	+
Opposes thumb to ...	+ (fingertips)	+ (index/little)	+ (index)	+ (index)	+
Extends wrist and fingers		+	--	*	--
Lower Extremity					
Maintains half-crooklying position			+	+	--
Flexes hip and knee while lying supine	+	+	+	+	+
Abducts hip in sidelying				*	--
Extends knee in sitting	+	+	+	+	+
Flexes knee in sitting				*	+
Flexes hip in sitting		+	+	+	+
Dorsiflexes ankle in sitting	+	+	+	+	+
Plantarflexes ankle in sitting				*	+
Extends knee and dorsiflexes ankle				*	+
Dorsiflexes affected ankle in standing	+	+	+	+	+
Flexes affected knee in standing	+	+	+	+	+
Abducts affected hip in standing	+	+	+	+	+

KEY: + = ADDED or RETAINED
– = DELETED
* = ADDED for further testing

(cont...)

TABLE 4.1 (cont)

The Evolution of STREAM Items

Item Description	Original STREAM	Revision 1 First Panel	Revision 2 Second Panel	Revision 3 Test Version	30 Item STREAM
BASIC MOBILITY					
Supine					
Moves knees side to side in crook lying				*	--
Bridges		+	+	+	+
Rolls		+ (both sides)	+ (either side)	+ (either side)	+
Moves from lying to sitting		+ (either side)	+ (either side)	+ (either side)	+
Sitting					
Maintains erect sitting position ...	+ (1 minute)	--	+ (20 counts)	+ (20 counts)	--
Raises folded arms to forehead	+	--	--	--	--
Moves folded arms side to side	+	--	--	--	--
Bends forward to touch floor	+	+	+	+	--
Standing					
Rises to standing from sitting	+	+	+	+	+
Maintains standing ...	+ (1 minute)	--	+ (20 counts)	+ (20 counts)	+
Places affected foot onto first step				*	+
Slides unaffected leg forward/backward	+	--	--	--	--
Lifts unaffected foot	+	--	--	--	--
Walking					
Turns 90^0 to unaffected side	+	+	+	+	--
Turns 90^0 to affected side	+	+	+	+	--
Takes 3 steps forwards	+	+	+	+	--
Takes 3 steps backwards	+	+	+	+	+
Takes 3 steps sideways to unaffected side	+	+	+	+	--
Takes 3 steps sideways to affected side	+	+	+	+	+
Walks 25' (10 meters)	+	+	+	+	+
Walks up 3 stairs not alternating feet	+	+	--	--	--
Walks down 3 stairs not alternating feet	+	+	--	--	--
Walks up 3 stairs alternating feet	+	+	+	+	--
Walks down 3 stairs alternating feet	+	+	+	+	+

KEY: + = ADDED or RETAINED
– = DELETED
* = ADDED for further testing

not been selected by the consensus panels were added to the test version of STREAM, as they were thought to warrant further testing. These additional items, denoted by an asterisk in Table 4.1, were either considered at an earlier stage of the instruments development, borrowed from other instruments, or generated by the investigators. They included primarily very low level or very high level movements. The items were added because, although the consensus panels had expressed a need for items in these extremes, none had been identified. Furthermore, it was felt that the analysis of internal consistency would identify any of these items that did not contribute to the content. Interestingly, five of these items performed very well in terms of the internal consistency analysis, met all the criteria for item retention, and thus are included in the completed instrument.

Table 4.2 summarizes the distributions of items included in each subscale of the STREAM at each stage of revision. There were minor fluctuations in the number of items in each subscale at the first and second revisions, yet the total number of items remained constant at thirty-four items. The test version of the STREAM included forty-three items that were relatively balanced amongst the subscales, with twelve items assessing lower extremity movement, thirteen items assessing upper extremity movement, and eighteen items assessing basic mobility. Finally, the completed instrument included a total of thirty items that were distributed evenly between each of the three subscales. Details of the process of item reduction, which led to the final version of STREAM, are provided in section 4.4.

TABLE 4.2

Subscale Evolution: number of items in the subscales following each revision

Subscale	Original STREAM	Revision 1 First Panel	Revision 2 Second Panel	Revision 3 Test Version	30 Item STREAM
Limb Movement:	15	19	18	25	20
upper extremity	9	12	10	13	10
lower extremity	6	7	8	12	10
Mobility	19	15	16	18	10
Total # of items	**34**	**34**	**34**	**43**	**30**

4.1.3 The evolution of STREAM scoring

During the first consensus panel a number of possible scoring schemes had been suggested (shown in Appendix 3.2.4), and these were presented to the second panel for further deliberation. As consensus had still not been fully achieved regarding the wording of the scoring instructions, a follow-up questionnaire for a revised scoring scheme (Appendix 3.2.6) was sent to the twenty panel participants to verify the acceptability of this revised scoring. Responses to the questionnaire were obtained from eighteen of the panelists (two therapists were away on vacation); seven therapists rated the scoring scheme acceptable as presented, nine said it would be acceptable with minor changes, and two felt that it required 'major' changes, although the comments from these therapists were identical to the comments of therapists suggesting that minor changes were needed.

The majority of panel participants expressed confusion over the proposed scoring of category '**1**' (Appendix 3.2.6), and in particular '**1c**'. Category '**1**' was intended to cover any movements between the two extremes of being completely unable (score **0**) or able to perform the task (score **2** for limb movement, **2** or **3** for mobility). In the proposed scoring scheme, category '**1**' was divided into the following three sub-categories:

1a: movements performed only *partially*,
1b: movements performed *completely*, but with an *abnormal* pattern, and
1c: movements performed *partially* with a *normal* pattern <u>or</u> *completely*, but in an *abnormal* pattern.

Due to the confusion expressed by the respondents, the format for scoring category '**1**' was revised to the present format, including the following three sub-categories:

1a: movements performed only *partially* and in an *abnormal* pattern,
1b: movements performed *partially*, but with a *normal* pattern, and
1c: movements performed *completely*, but with an *abnormal* pattern.

In addition to the confusion over scoring category '**1**', the panel participants suggested some minor changes in the wording of the scoring scheme. The majority of these changes were incorporated in the final version of the scoring scheme, shown in Figure 4.1. To further clarify the wording of the scoring scheme, the Glossary of Scoring Terms was developed, and is included in the test manual (Appendix 8.3). In addition, the rationale for the scoring format is explained in the test manual, and is presented in detail in section 5.1.3.

FIGURE 4.1
Scoring of the Revised STREAM
I. VOLUNTARY MOVEMENT OF THE LIMBS

0	**unable** to perform the test movement through any appreciable range (includes flicker or slight movement)
1 a.	able to perform only **part** of the movement, and with **marked deviation** from normal pattern
b.	able to perform only **part** of the movement, but in a manner that is **comparable to the unaffected side**
c.	able to **complete** the movement, but only with **marked deviation** from normal pattern
2	able to **complete** the movement in a manner that is **comparable to the unaffected side**
X	*activity not tested (specify why* **ROM, Pain, Other (reason)***)*

II. BASIC MOBILITY

0	**unable** to perform the test activity through any appreciable range (ie. minimal active participation)
1 a.	able to perform only **part** of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, and with **marked deviation** from normal pattern
b.	able to perform only **part** of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, but with a grossly **normal** movement pattern
c.	able to **complete** the activity independently, with or without an aid, but only with **marked deviation** from normal pattern
2	able to **complete** the activity independently with a grossly **normal** movement pattern, but **requires an aid**
3	able to **complete** the activity independently with a grossly **normal** movement pattern, **without an aid**
X	*activity not tested (specify why.* **ROM, Pain, Other (reason)***)*

In summary, in the final version of the scoring for STREAM a simple three point ordinal scale is employed for scoring voluntary movement of the limbs; a patient is scored '**0**' if **unable** to perform the test movement, '**1**' if a movement can only be **partially** completed or if the movement is performed **with marked deviation** from a normal pattern (as compared to the unaffected limb where possible), and '**2**' if **able to complete** the movement in a manner that is qualitatively and quantitatively near normal. The same scoring scheme is employed for the basic mobility subscale, except that an additional category is included to allow for independence with the help of an aid. If a test item cannot be performed due to pain or limited passive range, or if a movement or activity is limited because of other reasons (such as perceptual/cognitive deficits, amputation, impaired vision, etc.) then this is indicated by scoring an '**X**' and indicating the reason. When item(s) are scored as '**X**', the subscale and total scores are transformed so that they are out of 100. A subscale scoring form has been devised to

assist in summing item scores and transforming scores to scores out of 100, and is included in Appendix 8.3.

Table 4.3 was designed by therapists who were pilot testing the instrument, to facilitate scoring; it summarizes the revised scoring in a two-by-three table of quality of the movement in relation to amplitude of active movement.

TABLE 4.3
Scoring STREAM: the relationship between quality and amplitude of active movement

		AMPLITUDE OF ACTIVE MOVEMENT		
		None	Partial	Complete
QUALITY OF MOVEMENT	Marked Deviation	0	1 a	1 c
	Grossly Normal	0	1 b	2 (3)*

Note: * For the mobility subscale, movements completed independently with grossly normal pattern using an aid are scored as '2', without an aid are scored as '3'

4.2 Internal Consistency of the Test Version of STREAM

The characteristics of the subjects whose scores on STREAM were used in the initial evaluation of internal consistency of the test version of STREAM (and for evaluating reliability of scoring of individual items), and a description of the raters that contributed the data, are presented in section 4.6 (Direct Observation Reliability Study). The results of the initial evaluation are, however, presented here, as they were used in the second stage of refining the content of STREAM for item reduction.

The evaluation of internal consistency of the test version of STREAM involved computing inter-item, item to subscale and item to total score correlations, and alpha coefficients for subscales and for the STREAM as a whole. For each of the forty-three items, correlations with each of the other forty-two items, and with subscale and total STREAM scores were computed. A correlation matrix of 1764 coefficients (42 x 42 items) was produced. The inter-item, item to STREAM total, and item to subscale total score correlations are presented in Appendix 9.2. Inter-item correlations ranged from 0 to 1.0, with 8.2 % (145 / 1764) of the correlations falling below 0.4, 72.8 % (1284 / 1764) between 0.4 and 0.85, and 19.0 % (335 / 1764) above 0.85. The individual item to total correlations ranged from 0.49 to 0.95. Item to subscale

correlations ranged from 0.47 to 0.99. The alpha coefficients, calculated for each of the forty-three items to determine the effect that omitting that particular item would have on the overall alpha, ranged from 0.987 to 0.988; alphas for the STREAM subscales were 0.972 for the lower extremity, 0.979 for mobility, and 0.985 for the upper extremity; the overall alpha coefficient for the forty-three item instrument was 0.988.

4.3 Kappa Statistics for Items on the Test Version of STREAM

For each of the forty-three items included in the test version of STREAM, the indices of crude agreement and expected agreement using quadratic weights are presented in Table 4.4, along with the kappa statistics derived from these two indices. The 95% confidence intervals associated with each quadratically weighted kappa statistic are also given.

Over the forty-one items where kappa could be calculated, the coefficients ranged from 0.32 to 1.0. This distribution is summarized in Figure 4.2 as a stem and leaf plot. Clearly, the kappas of 0.8 and 0.9 predominated, indicating excellent agreement. Of the four items with less than excellent agreement, two items (item 6: hip abduction in side lying, and item 7: moving knees side to side) had been added to the test version of STREAM, after the second consensus panel. Also, there was very little variability of scoring (shown in Appendix 9.1) for two items (item 10: maintaining erect sitting, and item 28: reaching to the floor) as both of these items were performed well by all the patients assessed in this study.

FIGURE 4.2

Stem and Leaf Plot Illustrating the Distribution of Quadratically Weighted Kappas from the Direct Observation Reliability Study of Inter-Rater Agreement

POOR	**0.1**	
AGREEMENT	**0.2**	
(< .4)	**0.3**	2
	0.4	
MODERATE	**0.5**	3 8
AGREEMENT	**0.6**	5
(.4 -- .75)	**0.7**	
EXCELLENT	**0.8**	1 1 6 6 6 6 7 7 8 9 9
AGREEMENT	**0.9**	0 0 1 2 2 3 4 4 4 4 5 5 5 5 5 6 6 6 6 7 7 77
(> .75)	**1.0**	0 0 0

Note: Kappas given for 41 items for which kappa was defined.

TABLE 4.4

Indices of Agreement for Pairs of Raters Scoring Twenty Subjects

Item Number	Index of Crude Agreement (%)	Expected[a] Agreement (%)	Weighted Kappa (K_W) [b]	95% Confidence Interval for K_W
Upper Extremity Subscale:				
1	84	65	.89	0.76--1.02
2	95	68	.96	0.88--1.04
11	85	63	.90	0.79--1.01
12	85	69	.88	0.75--1.01
13	79	61	.86	0.73--1.00
14	85	64	.90	0.78--1.01
15	80	65	.86	0.71--1.00
16	100	67	1.00	1.00--1.00
17	100	62	1.00	1.00--1.00
18	90	66	.93	0.82--1.03
19	90	58	.94	0.86--1.02
20	95	61	.97	0.90--1.03
21	94	68	.96	0.87--1.04
Lower Extremity Subscale:				
3	85	95	.32	- 0.27--0.90
4	90	73	.91	0.77--1.04
6	74	86	.53	0.20--0.85
22	95	78	.94	0.83--1.06
23	90	80	.87	0.71--1.04
24	100	74	1.00	1.00--1.00
25	94	76	.94	0.82--1.06
26	85	65	.89	0.77--1.02
27	95	81	.94	0.82--1.05
31	95	76	.95	0.85--1.05
32	74	65	.81	0.66--0.97
33	95	60	.96	0.89--1.04

(cont...)

a Expected agreement assuming quadratic weights.

b Quadratically weighted Kappa statistic.

TABLE 4.4 (cont)

Item Number	Index of Crude Agreement (%)	Expected[a] Agreement (%)	Weighted Kappa (K_W) [b]	95% Confidence Interval for K_W
Basic Mobility Subscale:				
5	90	79	.86	0.65--1.08
7	80	87	.58	0.16--1.00
8	95	84	.86	0.59--1.13
9	95	80	.97	0.92--1.03
10	95	98	--[c]	--
28	100	100	--[c]	--
29	70	81	.65	0.34--0.95
30	85	82	.81	0.56--1.06
34	90	74	.96	0.90--1.02
35	84	78	.92	0.83--1.01
36	89	77	.87	0.67--1.07
37	89	77	.95	0.87--1.02
38	89	73	.95	0.89--1.02
39	84	77	.92	0.83--1.01
40	89	77	.95	0.88--1.02
41	89	78	.95	0.87--1.03
42	95	78	.97	0.92--1.03
43	95	78	.97	0.92--1.03

a Expected agreement assuming quadratic weights.
b Quadratically weighted Kappa statistic.
c Kappa undefined because either the numerator or denominator was 0.

4.4 Item Reduction

The specific criteria used for the process of item reduction were presented in Table 3.1 and section 3.1.4. In brief, two or more of the following characteristics would result in an item being eliminated: a preponderance of very high or very low inter-item correlations (more than 1/3 (14/42) of the inter-item correlations falling outside the range of 0.4--0.85), very high correlation(s) ($r > 0.9$) with item(s) measuring the same limb segment or a closely related mobility task, a very high or very low item-to-total correlation (outside the range of 0.4--0.85), an item Kappa statistic below 0.6, questionable face validity or clinical relevance. Thus, the decision to keep or omit items on the completed instrument was based on a combination of information, including the data obtained in the analysis of internal consistency of the test version of STREAM, item kappa statistics derived from the direct observation reliability study

(presented in Table 4.4), and judgements made by the consensus panels regarding the importance of the item in measuring the construct of motor recovery. In addition, insight regarding the clinical relevance and practical limitations of the items gained during the direct observation reliability study contributed to the decision process. The items that were retained in the final version of STREAM are noted in Table 4.5, along with the number of inter-item correlations that fell outside the acceptable range of 0.4 to 0.85, the item to total correlations, and a comment regarding the item's clinical relevance. Similarly, Table 4.6 presents the items that were excluded from the final version of STREAM, and the rationale on which these decisions were based.

TABLE 4 .5
STREAM Items Retained

Item Number	Out of Range Item-to-item Correlations [a]	Item-to-total Correlation	Kappa Statistic [b]	Clinical Relevance
Upper Extremity Subscale:				
1 protract scapula**	4	.88*	.89	unique item; low level
2 extend elbow **	1	.76	.96	unique item
11 shrug shoulders	7	.89*	.90	unique item; low level
12 touch head	12	.81	.88	functional
13 hand to sacrum	9	.73	.86	unique item; low level
15 full elevation	8	.81	.86	unique item
17 supinate/pronate	12	.78	1.00	unique item
18 close hand	8	.74	.93	unique item
19 open hand	13	.78	.94	unique item; higher level
20 opposition	13	.72	.97	unique item; high level
Lower Extremity Subscale:				
4 supine hip flexion	5	.73	.91	unique item; functional; low level
22 sitting hip flexion	4	.85*	.94	unique item; functional; low level
23 sitting knee ext	2	.86*	.87	low level; functional
24 sitting knee flex**	9	.92*	1.00	unique item; functional

Note: * = unacceptable value (see criterion for item reduction, Table 3 1). **(cont...)**
** = added for statistical verification of content (not contributed by panels)
a Number of inter-item correlations outside of range 0.4--0.85, out of a total of 42 .
b Quadratically weighted Kappa statistics from direct observation reliability study.

TABLE 4.5 (cont)
STREAM Items Retained

Item Number	Out of Range Item-to-item Correlations [a]	Item-to-total Correlation	Kappa Statistic [b]	Clinical Relevance
25 sitting dorsiflexion	6	.90*	.94	unique item; higher level
26 plantarflexion**	11	.86*	.87	unique item; functional; higher level
27 extends knee and dorsiflexes**	7	.89*	.94	high level; selective movement
31 standing hip abd	12	.87*	.95	unique item; high level
32 standing knee flexion	8	.82	.81	high level; isolated movement
33 standing dorsiflexion	7	.87*	.96	high level; isolated movement
Basic Mobility Subscale:				
5 rolling	2	.63	.86	unique item; functional; low level
8 bridging	10	.60	.86	unique item; functional; low level
9 supine to sitting	7	.82	.95	unique item; functional; higher level
29 rises to standing	0	.79	.65	unique item; functional; higher level
30 maintains standing	1	.82	.81	unique item; functional; low level standing activity
34 step onto step **	11	.90*	.96	unique item; functional; higher level
36 3 steps backward	9	.89*	.87	unique item; functional; higher level
38 3 steps to affected side	13	.94*	.95	unique item; functional; higher level
41 walks 10 m	5	.82	.95	unique item; functional; high level
43 down 3 stairs	9	.91*	.97	unique item; functional; high level

Note: * = unacceptable value (see criterion for item reduction, Table 3.1).
** = added for statistical verification of content (not contributed by panels)

a Number of inter-item correlations outside of range 0.4--0.85, out of a total of 42 .

b Quadratically weighted Kappa statistics from direct observation reliability study.

TABLE 4.6
STREAM Items Eliminated

Item Number	Out of Range Item-to-item Correlations [a]	Item-to-total Correlation	Kappa Statistic [b]	Clinical Relevance
Upper Extremity Subscale:				
14 90° shoulder flexion	10	.87*	.90	incorporated in full flexion (r=.93)
16 90° shoulder abduction	11	.86*	1.00	incorporated in hand to head (r=.91)
21 extends wrist and fingers**	9	.85*	.96	may be more limited by ROM than motor recovery; r=.92 with opens hand & opposition
Lower Extremity Subscale:				
3 maintains 1/2 crooklying	8	.54	.32*	weak association with subscale (r=.5 with subscale total)
6 sidelying hip abduction**	2	.73	.53*	may be limited by strength (bilat weakness; compensate with flexion); more functional in standing
Basic Mobility Subscale:				
7 rotates knees side to side**	14*	.59	.58*	incorporated in rolling / supine to sit; weak association with subscale (r=.6)
10 maintains sitting	17*	.51	--	no variability; weak (r=.57) association with subscale
28 reaches to floor	21*	.49	--	no variability; weak (r=.47) association with subscale; may reflect general health and flexibility vs motor recovery
35 3 steps forward	12	.92*	.92	incorporated in 10m walk (r=.87); r=.96 with 3 steps backward

Note: * = unacceptable value (see criterion for item reduction, Table 3.1). **(cont...)**

** = added for statistical verification of content (not contributed by panels)

a Number of inter-item correlations outside of range 0.4--0.85, out of a total of 42.

b Quadratically weighted Kappa statistics from direct observation reliability study.

TABLE 4.6 (cont)
STREAM Items Eliminated

Item Number		Out of Range Item-to-item Correlations [a]	Item-to-total Correlation	Kappa Statistic [b]	Clinical Relevance
Basic Mobility Subscale (cont):					
37	3 steps to unaffected side	14*	.95*	.95	r=.97 with 3 steps to aff; towards unaff side may be initiated by compensation vs indicative of motor recovery
39	turns 90° to unaffected side	15*	.95*	.92	r=.99 with 90° to aff; towards unaff side may be initiated by compensation vs indicative of motor recovery
40	turns 90° to affected side	12	.94*	.95	r=.99 with 3 steps to aff side
42	up 3 stairs	9	.91*	.97	r=1.00 with up 3 steps

Note: * = unacceptable value (see criterion for item reduction, Table 3.1).
** = added for statistical verification of content (not contributed by panels)
a Number of inter-item correlations outside of range 0.4--0.85, out of a total of 42 .
b Quadratically weighted Kappa statistics from direct observation reliability study.

4.5 Results of the Direct Observation Reliability Study

4.5.1 Characteristics of the subjects

Table 4.7 summarizes the characteristics of the twenty subjects who participated in the reliability study at the JRH, and the six additional low level subjects whose scores on STREAM were used in the internal consistency analysis. To elaborate, of the twenty subjects at the JRH, four had had previous CVAs involving the same side; six subjects were aphasic; three subjects had prominent perceptual and memory impairments; two subjects complained of shoulder pain; two subjects spoke neither English nor French, and were assessed with the help of a family member acting as translator. Interestingly, the majority (14/20) of subjects had left hemisphere CVAs, probably reflecting a greater propensity to admit patients needing a multitude of rehabilitation services. Of the six additional low level subjects, three subjects were aphasic and four of the subjects had prominent perceptual and cognitive impairments. The total scores on the test version of STREAM (raw score out of a maximum of 104, transformed to be out of 100) for these six low level subjects were 7, 7, 19, 20, 28, and 33.

TABLE 4.7

Characteristics of the Subjects Participating in the Direct Observation Study

Type of CVA	Side of CVA	Sex	Age	Days Post CVA	Comments / Comorbid Conditions
Ischemic	R	M	60	77	Shoulder Pain
Ischemic	R	M	67	141	1 Previous CVA (same side)
Ischemic	R	M	70	47	3 Previous CVAs (same side)
Ischemic	R	M	86	130	Language Barrier
Ischemic	R	F	73	119	Mild Perceptual/Memory Deficits
Ischemic	L	M	61	110	Aphasic/Perceptual Problems
Ischemic	L	M	67	66	Expressive Aphasia
Ischemic	L	M	70	70	Aphasic
Ischemic	L	M	75	105	Aphasic/Language Barrier
Ischemic	L	F	47	>1460	2 Previous CVAs, Lupus
Ischemic	L	F	48	113	Mild Aphasia
Ischemic	L	F	63	92	Mild Ataxia U/E
Ischemic	L	F	69	94	1 Previous CVA (full recovery)
Ischemic	L	F	71	122	Shoulder Pain
Ischemic	L	F	79	64	Hip Fracture (pinned-WBAT)
Ischemic	L	F	80	52	Ataxia & Vertigo; OA Knees
Hemorrhagic	R	F	50	155	Aneurysm Clipped, Lobectomy
Hemorrhagic	L	M	58	89	Expressive Aphasia
Hemorrhagic	L	M	59	101	No Complications
Hemorrhagic	L	M	80	238	Perception/Memory Impaired
Additional Low Level Subjects (for internal consistency analysis)					
Ischemic	R	F	81	44	Perception/Cognition Impaired
Ischemic	L	M	63	15	Aphasic/Perceptual Problems
Ischemic	L	F	67	21	Mild Aphasia
Hemorrhagic	R	F	34	16	Migraines
Hemorrhagic	L	F	57	18	Perception/Cognition Impaired
Hemorrhagic	L	M	70	41	Perception/Cognition Impaired,Aphasia

4.5.2 Characteristics of the raters

The raters were experienced physiotherapists (mean 5 years; range 2 to 9 years) who had worked with stroke patients (mean 2.5 years; range 1.5 to 3.5 years). The six raters participating in the reliability study at the JRH included four therapists working at an in-patient rehabilitation setting (the JRH), one therapist from an acute care setting, and the author (KD). The clinical backgrounds of these therapists are presented in greater detail in Appendix 7.

4.5.3 Internal consistency of the completed STREAM

Using only the data for the thirty items that were retained, the individual item to total correlations for the completed STREAM were calculated for our sample of twenty-six. as had been done on the test version (presented in 4.2). These values, presented in Table 4.8, were comparable (albeit generally slightly lower, as is expected with fewer items included) to those for the forty-three item instrument, and ranged from 0.579 to 0.926. The alpha coefficients reflecting the effect that omitting a particular item would have on the overall alpha, ranged from 0.982 to 0.984; alphas for the STREAM subscales were 0.965 for mobility, and 0.979 for both of the limb subscales; the overall alpha coefficient for the thirty item instrument was 0.984.

TABLE 4.8

Correlations of Item to Subscale and Total Scores on the Thirty Item STREAM

	Item Number		Correlation with:	
Subscale	**Test Version**	**(Completed)**	**Subscale Score**	**Total STREAM Score**
Upper Extremity:	1	(1)	.815	.883
	2	(2)	.840	.768
	11	(7)	.852	.904
	12	(8)	.952	.833
	13	(9)	.905	.749
	15	(10)	.875	.818
	17	(11)	.967	.793
	18	(12)	.914	.752
	19	(13)	.947	.800
	20	(14)	.907	.736
Lower Extremity:	4	(3)	.852	.740
	22	(15)	.864	.846
	23	(16)	.878	.856
	24	(17)	.918	.915
	25	(18)	.895	.897
	26	(19)	.960	.861
	27	(20)	.937	.886
	31	(23)	.895	.851
	32	(24)	.926	.817
	33	(25)	.897	.876
Basic Mobility:	5	(4)	.610	.636
	8	(5)	.585	.579
	9	(6)	.890	.798
	29	(21)	.827	.772
	30	(22)	.838	.795
	34	(26)	.954	.880
	36	(27)	.923	.875
	38	(28)	.972	.926
	41	(29)	.851	.813
	43	(30)	.938	.892

4.5.4 Inter-rater agreement for individual items on the STREAM

Table 4.9 shows, for each item, the distribution of perfect agreement, disagreement by one category, and disagreement by two categories. The table is divided into three parts. The first part lists the items related to the upper extremity subscale; followed by the items pertaining to the lower extremity and basic mobility subscales. The number in the last column indicates the number of subjects scored. Where the number of ratings was less than twenty, the item could not be scored for some subjects due to restricted range, pain, or other limitations. Although not presented, in fourteen of fifteen instances, the raters concurred not only that the item could not be scored, but also on the reason why it could not be scored.

For each subscale, the total number of ratings is given: 197 for the upper extremity, 193 for the lower extremity, and 195 for the basic mobility subscales respectively. At the foot of the table appears the proportion of ratings where the agreement was perfect or where there was disagreement. For example, in the upper extremity subscale, a total of 174 (54+36+84) ratings, or 88.3% demonstrated perfect agreement. Similarly, disagreement by one category occurred in 11.7% (23/197) of ratings, and there were no disagreements of two categories for this subscale. Over all items, there were a total of 585 ratings (~20 x 30); perfect agreement occurred in 89.4% of ratings; disagreement by one category occurred for 9.6% of ratings: and, for only 1.0% were there disagreements of two categories.

Kappa statistics reflecting the agreement between raters for the scoring of STREAM items were presented in Table 4.4, section 4.3. Of the thirty items that were retained on the completed instrument, all had excellent agreement, with the exception of item 29 (rising from sitting to standing) which demonstrated only moderate agreement (Kappa was 0.65). In addition, none of the signed ranks for individual item scores (not presented) were significant (ie. the probabilities of obtaining these statistics, under the null hypothesis of no median difference in scores, were all $>.05$), indicating that there was good agreement between raters.

4.5.5 Inter-rater agreement within scoring category '1'

An additional focus of this project was to determine whether we could incorporate a classification representing quality of movement into a scoring system. To this end, category **'1'** has three divisions: **a:** partial amplitude of active movement, with marked deviation from normal pattern, **b:** partial amplitude of active movement, in a manner that is comparable to the unaffected side, and **c:** complete amplitude of active

movement, but with marked deviation from normal pattern. Table 4.10 presents the distribution of agreement and disagreement within category '1'. This table is cast similarly to Table 4.9. The three subscales are indicated, and the items related to each of these subscales are clustered under the appropriate sub-headings, as are the total number of ratings for each subscale. Only items receiving a score of '1' by both raters are described in Table 4.10.[2] For example, for item one, only three of the twenty subjects involved received ratings of '1' by both raters; the range over all items was from one (items 13 and 33) to ten (item 27) ratings within category '1'.

Overall, perfect agreement within category '1' scoring predominated (82.1% of all item ratings). For the upper and lower extremity subscales, the preponderance of disagreements were between '1a' and '1b' (disagreement related to quality of movement), and between '1a' and '1c' (disagreement related to quantity of movement). In contrast, for the basic mobility subscale, disagreement was relatively equally distributed over the three possibilities, including disagreements between '1b' and '1c' (disagreement related to both quality and quantity of movement).

The prevalence of ratings within each of the categories '1a', '1b' and '1c' can be derived from Table 4.10. The prevalences of scores obtained for many of the items were low, and resulted in correspondingly low kappa statistics. Kappa statistics could not be computed for three items because the denominator was zero (that is, expected agreement was equal to 1). Figure 4.3 is a stem and leaf plot of the kappa statistics obtained for the scores within category '1' on each item. Over all thirty items, with pairs of raters scoring '1', kappas ranged from undefined to 1.0. There was a wider range of kappas for the agreement on the scoring within category '1', than was found for the agreement between categories '0', '1', and '2' (and '3' for the mobility subscale). Values of the crude and expected agreement, and unweighted kappa for the agreement on the scoring within category '1' for each item are found in Appendix 9.4.

---

2 Disagreements between category '1' and any other category are reflected in Table 4.9.

TABLE 4.9

Distribution of Agreement for Pairs of Raters Scoring Twenty Subjects

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1 or 1,0	1,2 or 2,1	0,2 or 2,0	N*
Upper Extremity:							
1	5	3	8	0	3	0	19
2	4	4	11	0	1	0	20
11	5	3	9	2	1	0	20
12	5	6	6	1	2	0	20
13	5	1	9	1	3	0	19
15	5	4	7	2	2	0	20
17	6	4	9	0	0	0	19
18	5	5	8	1	1	0	20
19	7	2	9	1	1	0	20
20	7	4	8	0	1	0	20
TOTAL	54	36	84	8	15	0	197
		88.3 %			11.7 %		
Lower Extremity:							
4	3	5	10	0	2	0	20
22	2	8	9	0	1	0	20
23	1	8	9	1	1	0	20
24	3	7	10	0	0	0	20
25	2	4	11	0	1	0	18
26	6	4	7	0	3	0	20
27	1	9	9	1	0	0	20
31	2	7	9	1	0	0	19
32	4	3	7	4	1	0	19
33	5	1	10	0	1	0	17
TOTAL	29	56	91	7	10	0	193
		91.2 %			8.8 %		

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	N*
Basic Mobility:										
5	0	7	0	11	0	0	1	0	1	20
8	0	4	1	14	0	0	0	0	1	20
9	0	8	1	10	0	0	1	0	0	20
29	0	7	1	6	0	4	0	0	2	20
30	0	5	2	10	0	1	1	0	1	20
34	2	8	0	8	0	2	0	0	0	20
36	1	8	1	7	0	1	0	0	1	19
38	2	5	1	8	0	2	0	0	0	18
41	2	9	1	5	0	2	0	0	0	19
43	3	7	4	4	0	1	0	0	0	19
TOTAL:	10	68	12	83	0	13	3	0	6	195
		88.7 %				8.2 %			3.1 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented

* Paired ratings on 20 subjects (due to missing data, the number of pairs of ratings range from 17 to 20).

TABLE 4.10

Distribution of Agreement on the Qualitative Divisions Within Category '1' [a]

Item	Perfect Agreement			Disagreement			N*
	1a	1b	1c	1a 1b	1b 1c	1a 1c	
Upper Extremity:							
1	1	1				1	3
2	2	2					4
11	1	1	1				3
12	1		2			1	4
13						1	1
15	2	1		1			4
17	1		1	1	1		4
18	1	1	2		1		5
19	1			1			2
20	1		3				4
TOTAL:	11	6	9	3	2	3	34
	76.5 %			8.8 %	5.9 %	8.8 %	
Lower Extremity:							
4	1		2			2	5
22	3		1	1		1	6
23	4	1	1		1	1	8
24	4	1		1	1		7
25	2	1		1			4
26	4						4
27	6	1	1	1		1	10
31	4		1	1			6
32		1	1	1			3
33			1				1
TOTAL	28	5	8	6	2	5	54
	75.9 %			11.1 %	3.7 %	9.3 %	
Basic Mobility:							
5	1		6				7
8	1	1			2		4
9	2		5		1		8
29	4		3				7
30			5				5
34	2		5	1			8
36	4		2			2	8
38	3		1	1			9
41	4		5				9
43	5		2				7
TOTAL:	26	1	34	2	3	2	68
	89.7 %			2.9 %	4.4 %	2.9 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

* The number of subjects (out of 20) scored '1' by both raters.

a Category one has three divisions: **1a: partial** movement with **marked deviation** from normal pattern
1b: partial movement with **grossly normal** pattern
1c: complete movement with **marked deviation** from normal pattern

FIGURE 4.3

Stem and Leaf Plot of Unweighted Kappas Showing Inter-Rater Agreement on the Qualitative Divisions Within Category '1'

	0.0	0 0 [a]
POOR	**0.1**	7 8
AGREEMENT	**0.2**	0 7
(< .4)	**0.3**	9
	0.4	0 4 7
MODERATE	**0.5**	0 0 0 0 8 8
AGREEMENT	**0.6**	9
(.4 -- .75)	**0.7**	3
	0.7	5 6
EXCELLENT	**0.8**	
AGREEMENT	**0.9**	
(> .75)	**1.0**	0 0 0 0 0 0 0

Note: Kappa was undefined for three items (denominator 0), therefore, 27 kappas are presented.
a Crude and expected agreement were equivalent (numerator 0) for two items.

4.5.6 Inter-rater agreement for subscale and STREAM total scores

Scores for the STREAM as a whole, and for each subscale, were calculated for each subject by summing the item scores (see Appendix 9.3). A total of fifteen items were scored as **'X'**, because seven of the patients could not perform them due to restricted range of motion, pain, or other reasons (eg. dizziness). Therefore, the scores have been transformed to scores out of 100. The procedure for transforming subscale and total scores is shown in Appendix 8.3.

The subscale and total STREAM scores given by the two raters for each of the twenty subjects are shown in Appendix 9.3 and in Figure 4.4. Figure 4.4 also shows the pattern of agreement between the two raters on subscale and total scores. The close proximity of the two lines indicates excellent inter-rater agreement that was consistent across the entire range of scores, with a maximum difference on total scores of twelve points between ratings. Interestingly, there was a tendency for the rater that was observing to score slightly higher than the rater doing the hands on assessment. The observers gave the same (seven subjects) or higher (ten subjects) total STREAM score than the assessors for seventeen of the twenty subjects (85%). To determine whether the difference between the scores given by the two raters was significant, signed ranks were computed for subscale and total scores. These test statistics, along with their related probabilities, are summarized in Table 4.11. None of the signed ranks were significant (at $p<.05$), indicating that there was good agreement between the raters.

The signed rank statistic was, however, nearly significant for the total scores, and this result corresponds with the trend of higher scoring by observers seen in Figure 4.4.

As well as signed rank statistics, Table 4.11 includes the GCCs for inter-rater agreement on the subscale and total scores for the twenty subjects. Three different models were used to calculate the GCCs: subjects and raters as random facets, subjects random and raters fixed, and subjects fixed with raters considered random facets. The GCC relating to each of these models indicates the extent to which we can generalize across raters (these particular raters, when raters are considered fixed facets; any raters, where raters are considered random facets) when subjects are the object of measurement. For example, for the upper extremity subscale, in order to differentiate between any subjects using any raters, the generalizability coefficient would be 0.994, indicating excellent generalizability across raters. The GCCs were identical for the subjects fixed and subjects as random facets models, where raters were considered random facets. This indicates that the raters contributed little variability to the scoring of STREAM; reliable STREAM scores should, therefore, be attainable by the general population of physical therapists, and not just by the selection of raters in this study. The GCCs were undefined when raters were considered fixed facets, as all of the variability was contributed by the subjects and the model was saturated. It can be seen from the GCCs obtained, that of the three subscales, the upper extremity subscale was the most reliably scored, followed by the lower extremity and basic mobility subscales.

TABLE 4.11

Summary of Inter-rater Agreement on Subscale and Total STREAM Scores for Rater Pairs Scoring Twenty Subjects

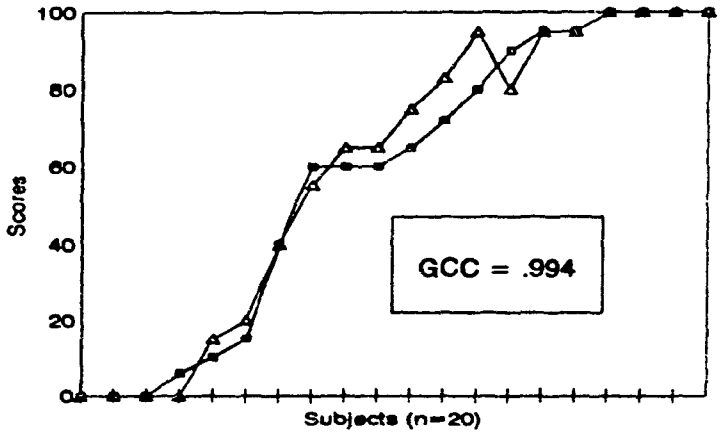
	Signed Rank	Generalizability [a] Coefficient
Upper Extremity:	-11 (.311) [b]	.994
Lower Extremity:	-8 (.281)	.993
Basic Mobility:	-16 (.172)	.982
TOTAL SCORES:	-26 (.072)	.995

a GCCs were equivalent for raters and subjects random, and subjects fixed with raters random models.
b Probabilities in brackets.

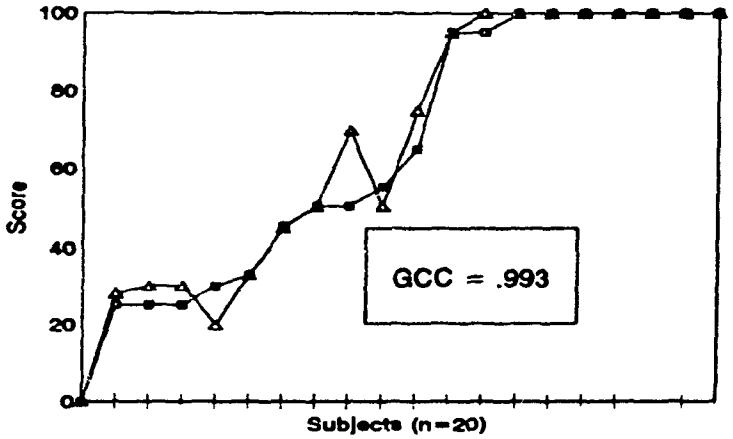
FIGURE 4.4

Direct Observation Study: Comparison of Two Ratings

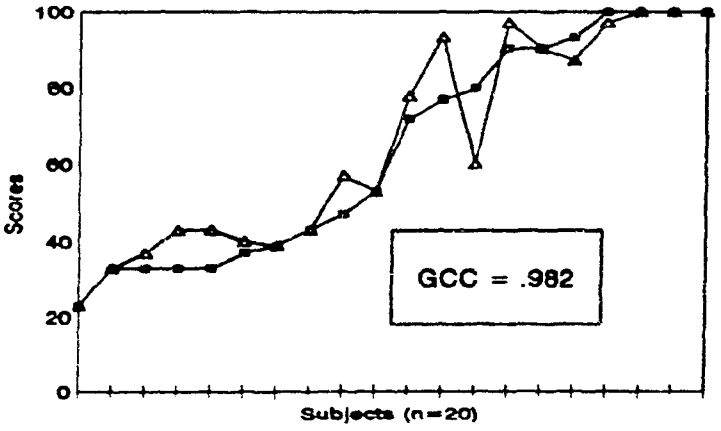
a) Upper Extremity Subscale Scores



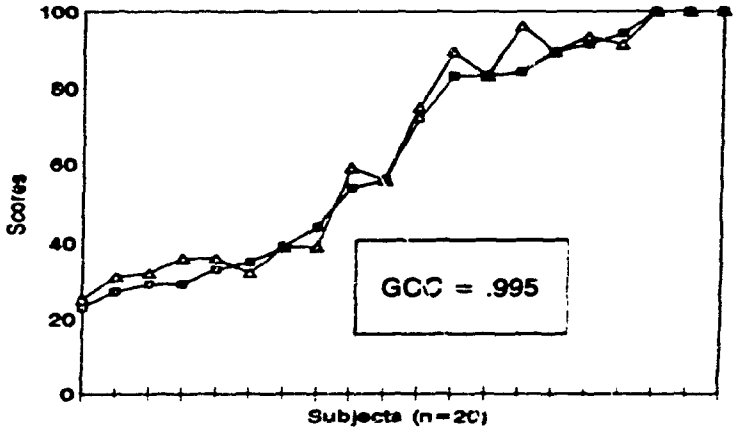
b) Lower Extremity Subscale Scores



c) Basic Mobility Subscale Scores



d) Total Scores



Note: scores have been transformed to scores out of 100
GCC = raters and subjects random model.

Key: □ Assessor △ Observer

4.6 Results of the Videotaped Assessments Reliability Study

4.6.1 Characteristics of the subjects

Table 4.12 summarizes the characteristics of the subjects who participated in the videotaped assessments reliability study. The subjects ranged from high to low functioning, and included one aphasic subject, two subjects with perceptual and memory impairments, and one subject with shoulder pain.

TABLE 4.12

Characteristics of Subjects Participating in the Videotaped Assessments Reliability Study

Type of Stroke	Side of CVA	Sex	Age	Days Post CVA	Comments
Ischemic	R	M	60	77	**High level: U/E impaired** Minor Shoulder Pain
Ischemic	L	M	61	110	**Moderate level motor function** Aphasic/Perceptual Problems
Hemorrhagic	L	M	80	238	**Moderate level motor function** Perceptual/Memory Impaired
Hemorrhagic	R	F	50	155	**Low level motor function** Aneurysm Clipped/Lobectomy

4.6.2 Characteristics of the raters

The twenty raters who participated in rating the tapes were all experienced as physiotherapists (mean 9 years; range 1 to 33 years), and had a minimum of six months of experience working with stroke (mean 4.5 years; range 0.5 to 11 years). They included therapists working in acute care, in- and out-patient rehabilitation, and LTC settings. Table 4.13 presents the clinical backgrounds of these raters.

4.6.3 Intra-rater agreement for individual items on the STREAM

Table 4.14 is organized in the same manner as Table 4.9 in the presentation of the results from the direct observation reliability study, except that, in Table 4.14 the distributions of ratings given in the two viewing sessions by each of the twenty raters (intra-rater agreement) are given. This is in contrast to the distribution of agreement

TABLE 4.13
Characteristics of Raters Participating in the Videotaped Assessments Reliability Study (N =20)

Service Area	Caseload	# of Stroke Patients Assessed in Past Year	Years of Experience Physio (Stroke)
Acute	Neuro	10-30	1 (0.5)
Acute	Mixed	10-30	1 (0.5)
Acute	Neuro	10-30	10 (5)
Acute	Neuro	10-30	3 (1)
Acute	Neuro	10-30	11.5 (2)
In-pt rehab/Acute	Neuro	>30	4 (4)
In-pt rehab/Acute	Mixed	10-30	5.5 (4)
In-pt rehab/Acute	Mixed	10-30	2 (1)
In-pt rehab	Geriatrics	10-30	5 (2)
In-pt rehab	Mixed	>30	20 (10)
In-pt rehab	Stroke	10-30	9 (4)
In-pt rehab	Stroke	10-30	8 (3)
In-pt rehab	Neuro	>30	7 (6.5)
In-pt rehab/LTC	Mixed	10-30	9.5 (3.5)
Out-pt rehab	Neuro	>30	24 (11)
Out-pt rehab	Neuro	>30	33 (11)
Out-pt ortho	Ortho	<10	6.5 (5)
LTC	Mixed	10-30	10 (7.5)
MSc Student	N/A	<10	4 (2.5)
MSc Student	N/A	<10	8 (4)

for rater pairs (inter-rater agreement) shown in Table 4.9. The distribution of agreement in each of the three categories (perfect agreement, disagreement by one category, and disagreement by two categories) for each item, and for each of the three subscales are presented. The number in the last column indicates the total number of ratings given for the four videotapes combined (ie. 4 videotapes x 20 raters = 80 ratings). There were no missing item scores, hence there are a total of 80 ratings for each item, a total of 800 ratings for each subscale (10 items x 80 ratings), and a grand total of 2400 ratings over all 30 items. Over all items, perfect agreement occurred in 85.7% of the ratings, disagreement by one category occurred for 12.1%, and for only 2.2% were there disagreements of two categories.

Four tables (similar to Table 4.14) showing the distributions of intra-rater agreement achieved on item scores for each of the videotapes (20 ratings for each item) can be found in Appendix 10.1. The distributions of intra-rater agreements over each of the

TABLE 4.14

Distribution of Agreement for Twenty Raters Scoring Four Videotapes

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1 or 1,0	1,2 or 2,1	0,2 or 2,0	N*
Upper Extremity:							
1	17	45	6	5	7	0	80
2	0	42	18	0	20	0	80
11	17	24	15	18	6	0	80
12	24	51	1	3	1	0	80
13	25	49	1	4	1	0	80
15	23	47	2	5	3	0	80
17	2	72	1	5	0	0	80
18	14	47	4	12	3	0	80
19	42	11	4	15	8	0	80
20	22	47	2	6	3	0	80
TOTAL:	186	435	54	73	52	0	800
		84.4 %			15.6 %		
Lower Extremity:							
4	13	18	39	6	4	0	80
22	0	43	23	0	14	0	80
23	0	56	21	1	2	0	80
24	2	46	21	6	5	0	80
25	22	31	16	4	7	0	80
26	19	24	21	12	4	0	80
27	1	55	20	4	0	0	80
31	6	35	22	10	6	1	80
32	20	26	19	14	1	0	80
33	40	16	17	5	2	0	80
TOTAL:	123	350	219	62	45	1	800
		86.5 %			13.4 %	0.1 %	

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	N*
Basic Mobility:										
5	0	56	0	7	0	3	0	0	14	80
8	0	7	0	65	0	0	1	0	7	80
9	0	29	23	11	0	5	4	0	8	80
29	0	22	24	20	0	13	1	0	0	80
30	0	11	0	45	0	1	1	0	22	80
34	0	49	2	20	0	9	0	0	0	80
36	0	54	0	19	0	7	0	0	0	80
38	0	51	1	20	0	8	0	0	0	80
41	0	58	0	20	0	2	0	0	0	80
43	0	57	0	20	0	3	0	0	0	80
TOTAL:	0	394	50	247	0	51	7	0	51	800
		86.4 %				7.3 %			6.4 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

* Twenty raters scoring four videotapes on two occasions.

three subscales, and over all scale items (ie. on total scores), for each of the four tapes individually (20 ratings per item), and for the four tapes combined (80 ratings per item) are summarized in Table 4.15. For example, it can be seen from the first row of Table 4.15 that, for the Tape A upper extremity subscale scores, 84.0% (168/200) of ratings occurred in the perfect agreement category, 16.0% (32/200) of the ratings were disagreements by one category, and none of the paired ratings differed by two categories. The ranks of the extent of perfect agreement on each of the subscales and for total scores for the individual videotapes are given in the second column of Table 4.15; these ranks corresponded roughly with the subjects' functional levels relative to each of the subscales. The two lower ranking videotapes (Tapes C and D) were the subjects with left sided CVAs and perceptual problems.

The signed ranks, and accompanying probabilities, indicating intra-rater agreement on individual test items for each of the four videotapes, are presented in Appendix 10.4. There were only two items ('maintain standing' of tape C, and 'sit to stand' of tape D) with significant ($p<.05$) signed ranks, reflecting a significant trend to score the item higher on the second viewing session. Thus, with the exception of those two items, there was generally good agreement between the raters' scores on the two viewing occasions, over all items on the four separate tapes.

4.6.4 Intra-rater agreement within scoring category '1'

The distributions of intra-rater agreement and disagreement within category '1' (similar to Table 4.10) for the item scores for each of the four videotapes are found in Appendix 10.1. Table 4.16 is organized in the same fashion as Table 4.15; it summarizes the distributions of intra-rater agreement and disagreement within category '1' for subscale and total scores, for each of the four videotapes individually, and for the four videotapes combined. The numbers of paired ratings which occurred within category '1' for each subscale and for total scores for each videotape (from a maximum possible of 20 ratings x 10 items in each subscale, or from 20 ratings x the total of 30 items (ie. 600 ratings for total scores) per tape) are noted in the last column of this table. Tape A had the fewest paired ratings within category '1' (143, over the 30 items), as this subject was relatively high functioning. Tape B (our lowest functioning patient), also had fewer ratings within category '1' than did the tape C and Tape D subjects. For the subscales of the videotapes, the number of ratings within category '1' ranged from three (for the tape A lower extremity subscale) to 146 (on the tape D basic mobility subscale). No meaningful ranks could be given for the extent of perfect

TABLE 4.15

Summary of Intra-Rater Agreement for Twenty Raters Scoring Four Videotapes on Two Occasions

	Rank[a]	Perfect Agreement (%) 0,0 1,1 2,2	Disagreement (%) One Category 0,1 or 2,1	Two Categories 0,2 or 3,1	N*
Upper Extremity:					200
Tape A	3	168 (84.0)	32 (16.0)	0	
Tape B	4	184 (92.0)	16 (8.0)	0	
Tape C	2	162 (81.0)	38 (19.0)	0	
Tape D	1	161 (80.5)	39 (19.5)	0	
Combined		84.4%	15.6%	0	800
Lower Extremity:					200
Tape A	4	193 (96.5)	7 (3.5)	0	
Tape B	3	168 (84.0)	31 (15.5)	1(0.5)	
Tape C	1	164 (82.0)	36 (18.0)	0	
Tape D	2	167 (83.5)	33 (16.5)	0	
Combined		86.5%	13.4%	0.1%	800
Basic Mobility:					200
Tape A	4	189 (94.5)	0	11 (5.5)	
Tape B	2	174 (87.0)	13 (6.5)	13 (6.5)	
Tape C	1	154 (77.0)	30 (15.0)	16 (8.0)	
Tape D	3	176 (88.0)	13 (6.5)	11 (5.5)	
Combined		86.4%	7.3%	6.4%	800
TOTAL SCORES:					600
Tape A	4	550 (91.7)	39 (6.5)	11 (1.8)	
Tape B	3	526 (87.7)	60 (10.0)	14 (2.3)	
Tape C	1	480 (80.0)	104 (17.3)	16 (2.7)	
Tape D	2	504 (84.0)	85 (14.2)	11 (1.8)	
Combined		85.7%	12.1%	2.2%	2400

* 20 raters scoring 30 items (10 items for each subscale) on each videotape; 20 raters scoring 30 items on the 4 videotapes (combined).

a Ranked in order of extent of perfect agreement.

TABLE 4.16

Summary of Intra-Rater Agreement Within Scoring Category '1' for Twenty Raters Scoring Four Videotapes on Two Occasions

	Perfect Agreement (%)			Disagreement (%)			
	1a	1b	1c	1a 1b	1b 1c	1a 1c	N*
Upper Extremity:							
Tape A	102 (83.6)			11 (9.0)	7 (5.7)	2 (1.6)	122
Tape B	42 (84.0)			8 (16.0)	0	0	50
Tape C	103 (93.6)			6 (5.5)	0	1 (09)	110
Tape D	116 (92.8)			9 (7.2)	0	0	125
Combined	89.2%			8.4%	1.7%	0.7%	407
Lower Extremity:							
Tape A	2 (66.7)			1 (33.3)	0	0	3
Tape B	65 (65.0)			20 (20.0)	4 (4.0)	11(11.0)	100
Tape C	89 (63.6)			39 (27.9)	5 (3.6)	7 (5.0)	140
Tape D	94 (87.0)			14 (13.0)	0	0	108
Combined	71.2%			21.1%	2.6%	5.1%	351
Basic Mobility:							
Tape A	18 (100)			0	0	0	18
Tape B	134 (95.0)			0	1 (0.7)	6 (4.3)	141
Tape C	97 (91.5)			8 (7.5)	0	1 (0.9)	106
Tape D	137 (93.8)			6 (4.1)	0	3 (2.1)	146
Combined	93.9%			3.4%	0.2%	2.4%	411
TOTAL SCORES:							
Tape A	122 (85.3)			12 (8.4)	7 (4.9)	2 (1.4)	143
Tape B	241 (82.8)			28 (9.6)	5 (1.7)	17 (5.8)	291
Tape C	289 (81.2)			53 (14.9)	5 (1.4)	9 (2.5)	356
Tape D	347 (91.5)			29 (7.7)	0	3 (0.8)	379
Combined	85.5%			10.4%	1.5%	2.6%	1169

* The number of paired ratings within category '1' (out of a maximum of 600 per tape) for 20 raters scoring 30 items on two occasions.

agreement within scoring category one (as was done in Table 4.15) because there were only three paired ratings within category '1' for the tape A lower extremity subscale.

Clearly, the majority of ratings within category '1' demonstrated perfect agreement. The basic mobility subscale had the highest proportion of ratings with perfect agreement (93.9% over all 4 tapes), followed by the upper extremity (89.2%), and the lower extremity (71.2%) subscales. Disagreements of two categories occurred more frequently on the lower extremity subscale (5.1% over the 4 tapes), than for either the mobility (2.4%) or the upper extremity (0.7%) subscales.

4.6.5 Intra-rater agreement for subscale and STREAM total scores

Figure 4.5 shows the pattern of agreement on the total STREAM scores given by the twenty raters for each of the four videotapes on the two occasions. The close proximity of the two lines suggests excellent intra-rater agreement. The paired ratings were generally within a few points of each other; the greatest difference between ratings on the two occasions on any of the tapes was nine points. Because there were no items scored as 'X' in this study, total STREAM scores were not transformed to be out of 100, and the raw scores out of a maximum possible score of seventy (ie. a maximum of twenty points on each of the limb subscales and thirty points on the mobility subscale) are shown. Plots of subscale scores for each of the four videotapes are found in Appendix 10.2. Tables of subscale and total STREAM scores for each of the four videotapes are found in Appendix 10.3. From these tables, and in Figure 4.5, it can be seen that, although intra-rater agreement was excellent, the subscale and total STREAM scores given were consistently slightly higher (an average of two points higher for total scores, over the four tapes) for the second rating session.

To determine whether the trend to score higher on the second session was significant, signed ranks were computed for subscale and total scores for the four videotapes; these statistical tests of agreement are presented in Table 4.17. The signed ranks were significant ($p < .05$) for all subscales on tapes C and D, indicating that the trend to score higher on the second viewing session was significant for these tapes.

Table 4.17 also summarizes the GCCs for subscale and total scores for the four videotapes. The subscale GCCs were identical for the raters fixed and raters random models, and differed only in the third decimal place for the total score GCCs. Unlike the results of the direct observation study, where the upper extremity subscale was the most reliably score, the GCCs in this study were slightly higher for the lower extremity

and mobility subscales, followed by the upper extremity subscale. All raters demonstrated excellent intra-rater agreement on scoring the 4 videotapes, with GCCs for individual raters ranging from 0.982 to 0.999 (subjects and occasions fixed model).

TABLE 4.17

Summary of Intra-rater Agreement for Subscale and Total STREAM Scores for Twenty Raters Scoring Four Videotapes on Two Occasions

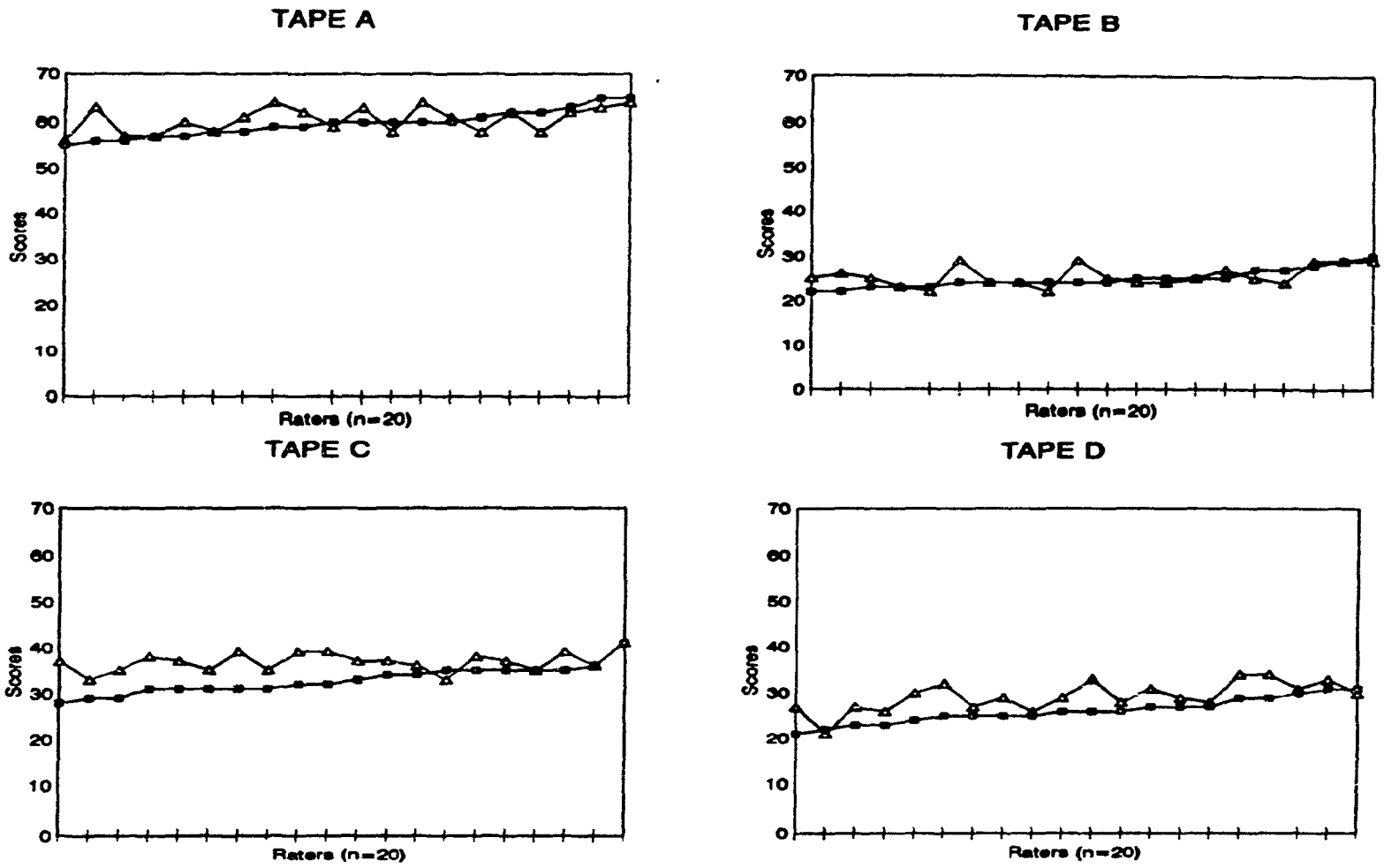
	Upper Extremity	Lower Extremity	Basic Mobility	TOTAL SCORES
Signed Rank:				
TAPE A	-6.5 (.801) [a]	-7 (.188)	-12.5 (.273)	-25.5 (.246)
TAPE B	-9.5 (.365)	6.5 (.659)	-13.5 (.427)	-17.5 (.076)
TAPE C	-35 (.049)	-39 (.012)	-83 (.0003)	-74.5 (.0001)
TAPE D	-53 (.009)	-26 (.045)	-33 (.001)	-99 (.0001)
Generalizability Coefficient [b]:				
	.963	.999	.999	.999

a Probabilities in brackets.
b Model = subjects and occasions fixed, raters random.

4.6.6 Inter-rater agreement for STREAM scores on the two occasions

For each of the two viewing sessions, inter-rater agreement on the scoring of the four videotapes was determined. That is, GCCs were computed for each occasion, providing two estimates of the agreement between the twenty raters on total STREAM scores given on the four taped assessments. The STREAM scores used to estimate inter-rater agreement for each of the two occasions are given in Appendix 10.3, and are shown in Figure 4.5. The relative flatness (slope near 0) of the lines that show the scores on the two occasions in Figure 4.5, indicates that the agreement between raters on the total STREAM scores was excellent on each occasion. On both occasions the GCCs were 0.999 (subjects fixed and raters random model).

FIGURE 4.5
Videotaped Assessments Reliability Study: Comparison of the Two Ratings on STREAM



Note: GCC = .999 (model = subjects and occasions fixed, raters random)

Key: □ Time 1 △ Time 2

4.7 Summary of the Results

Based on the consensus of the two panels of experts, whose members represented the potential consumers, the items and scoring of the original STREAM were revised to enhance content validity. The test version of STREAM was comprised of forty-three items, with three subscales of twenty-five limb movement items (scored on a 3-point scale) and eighteen basic mobility items (scored on a 4-point scale). Item reduction, based on preliminary estimates of reliability of scoring items and internal consistency of the test version of STREAM, led to the completed version of STREAM, comprised of thirty items equally distributed amongst the three subscales. The internal consistency of this thirty item instrument was excellent, with Cronbach's alphas of greater than 0.98 on subscales and overall. The reliability was also excellent, both within and between raters, with GCCs of 0.99 for total scores, and from 0.96 to 0.99 for subscale scores.

CHAPTER 5

DISCUSSION

5.0 Introduction

This study has contributed information on several of the measurement properties of STREAM, including: content validity, internal consistency, inter-rater and intra-rater reliability. In this chapter, the outcomes and implications of the content development and validation phase, the internal consistency analysis, and the direct observation and videotaped assessments reliability studies, are summarized and discussed. The two different methods used for evaluating the reliability of STREAM are contrasted, and are compared to testing done on other measures of motor function for stroke. Next, a general comparison is made between the STREAM and related measures of motor function. The limitations of this study, and the potential limitations of STREAM as a clinical measure of motor recovery are identified; implications for future research on STREAM are presented; and finally, the conclusions drawn from this study are posed.

5.1 Content Development and Validation

In the process of refining the STREAM, close attention was paid to the validity of the content and to the utility of the instrument. Our consensus panels deliberated over the selection of items, and produced the forty-three item test version of STREAM. This instrument was then pared down to include thirty items that were, based on an analysis of the internal consistency, all contributing information that was related to the instrument as a whole but not redundant. The items retained also demonstrated face validity and were reliably scored. Interestingly, the recommendations from the respondents of the content verification survey closely paralleled those of the consensus panels. This convergence of feedback from a large number of therapists (including sixty-four from the survey, and twenty from the panels) lends further support to the appropriateness of the items included.

In general, the changes which occurred in the evolution of STREAM from the original instrument, to the test, and finally to the completed versions involved primarily changing the wording of the original items to improve clarity, adding some very high and very low level items to eliminate ceiling and floor effects so that the STREAM could be used with the spectrum of motor impairments, and changing the scoring to

improve sensitivity. In addition, items were modified so that they were more functionally oriented, and a qualitative aspect was added to the scoring scheme.

5.1.1 Intent and scope of STREAM

In devising the STREAM, a number of fundamental properties were considered, including: the specific use to which it would be put, the domains to be included, how these domains would be evaluated, and the populations for which the measure should be suitable. In addition, our goal was to produce a user friendly instrument, based on the input of its potential consumers, that could be realistically incorporated into routine clinical practice.

The STREAM is intended to be used as an outcome measure for evaluating treatment effects, and for monitoring motor improvement. STREAM was not designed to be used in isolation as an assessment tool for diagnosis and treatment planning, which would require considerably more detailed information. None the less, we wanted to achieve an instrument that would provide a relatively comprehensive profile of a patient's ability to move. Two aspects or domains of movement are incorporated. That is, voluntary movements of the limbs are evaluated to provide a profile of global motor status, and basic mobility is evaluated to give a more functional picture of motor recovery reflecting the integration of movement of trunk and limbs. Limb movement items include movements in several directions at each joint, and mobility items include activities performed in lying, sitting standing and walking. Because we wanted the STREAM to reflect a patient's present level of motor functioning, actual performance of items is scored rather than relying on self (or therapist) reports. Since the STREAM is intended to be suitable for use at all levels of stroke rehabilitation, and for individuals with movement dysfunction ranging from mild to severe, a range of movements representing gradations of difficulty were included. And finally, the length and complexity of STREAM have been kept to a minimum to enhance its clinical utility, and, therefore, to make the STREAM more attractive for use in the clinical setting.

5.1.2 Item selection

A number of factors influenced the selection of items included in the STREAM. Our intention was to keep the STREAM as streamlined as possible, while still providing sufficient detail to adequately measure the construct of motor function for the spectrum of motor impairment in all phases of stroke rehabilitation. For discriminating between patients with differing levels of motor impairment, and for monitoring individual's

motor recovery over time, it was necessary to include movements of graded difficulty. Thus, the items included provide a relatively detailed profile of a patient's movement ability in terms of limb movements and basic mobility, and range from fairly rudimentary movements that can typically be performed early in the recovery process (eg. shrugging shoulders, rolling) to more complex selective movements (eg. opposition, stair climbing). Since the focus of STREAM is to measure the recovery of movement, and specifically motor ability, items were limited to simple movements and activities in an attempt to minimize the confounding effects that variables other than motor recovery, such as cognition, perception and communication, have on the measurement of motor status. For this reason, some of the mobility items (rolling and moving from supine to sitting) may be performed to either side, and only simple one or two step commands are used and may be supplemented by demonstration to facilitate communication.

The items also had to have acceptable psychometric properties, that is, they had to have face validity (ie. our two consensus panels had to consider the items to be important and representative for evaluating motor recovery), they had to be reliably scored, and they had to contribute information that was not already being contributed by other items (reflected by inter-item correlations). To ensure that the items included in the final version of STREAM would meet these standards, we evaluated the items on the test version of STREAM based on the criteria outlined in Table 3.1. To summarize the results of the item reduction (presented in section 4.4), a total of thirteen items were eliminated because they did not meet our criteria. Three upper extremity items (90^0 shoulder flexion, 90^0 shoulder abduction, and combined wrist/finger extension) were eliminated because they had very high correlations with other similar movements and with the total STREAM score. Two lower extremity items (half crooklying, and sidelying hip abduction) were eliminated because neither were reliably scored, and both had questionable clinical relevance--the first due to a weak association with the subscale, and the second because hip abduction is probably more functional when performed in standing than when performed in lying and because hospitalized elderly individuals often demonstrate difficulty with this movement bilaterally, regardless of diagnosis. Finally, eight mobility items were dropped, two of which (maintaining sitting and reaching to touch the floor) were eliminated on the basis of having no variability and no relationship to the other items on STREAM, and the remaining six on the basis of being redundant. Thus, the process of item reduction served to identify

and remove those items from the test version of STREAM that were not required or desirable.

5.1.3 The Scoring of STREAM

A variety of approaches have been employed by the existing published instruments for measuring the recovery of movement following stroke. For example, some instruments (Ashburn, 1982) have simply graded the movement in terms of quantity (eg. 'none / part / full' for limb activity; 'unable / with assistance / with aid / independent' for mobility), and have not addressed the issue of quality of movement at all. Others have ordered the test items hierarchically with respect to degree of presumed difficulty, based on the assumption that stroke patients follow a predictable or consistent pattern of recovery. With this Guttman-like scheme, the patient is evaluated on progressively more difficult tasks until unable to perform several consecutive items; the more items performed, the better the overall function and the higher the score of the individual (Lincoln and Leadbitter, 1979). Carr and Shepherd's MAS (Carr and Shepherd, 1985) utilizes a seven point ordinal scale, with each successive level of scoring representing a progressively higher level of function in terms of the quality of the movement. However, scoring of this instrument is relatively complicated because the seven ordinal levels are different for each item. The Fugl-Meyer (Fugl-Meyer et al, 1975) and Chedoke-McMaster (Gowland et al, 1991) scales have employed the concept of synergies to reflect the degree of recovery and quality of movement. There are several problems, however, in using the synergic approach to classify movements--1) the pattern of recovery of movement may not conform precisely to the rigid hierarchy, in which case scoring becomes difficult, and, 2) the movement patterns evaluated using this model often do not reflect everyday functional movement patterns that would perhaps be more meaningful indicators of motor recovery.

A number of scoring dimensions were incorporated in the scoring scheme of the revised STREAM, including scoring with respect to: active range or amplitude of voluntary movement possible, quality of movement, and the degree of assistance required (or level of independence). The amplitude of active movement was felt to be an important aspect to measure as it has obvious functional implications, and provides an objectively measurable reflection of changes in motor ability. From our content verification survey and consensus panels it was evident that clinicians wanted quality of movement documented. Therapists were particularly concerned with quality of movement in the early stages of recovery from stroke, with the emphasis shifting more to function as

motor recovery plateaued. Physical independence, specifically for carrying out mobility activities, was also felt to be an important indicator of motor recovery, and thus this dimension was incorporated into the scoring of the mobility subscale.

Since the scoring of STREAM was intended to be as simple as possible, we limited the categories for scoring the amplitude of movement to three ranges: no or minimal movement (ie. less than 10%of normal movement); complete movement (or at least 90% of full movement); and, any movement between these two extremes. These relatively wide scoring ranges enhance the STREAM's reliability, as clinicians are less likely to disagree among these choices than when more categories are involved. They are also a plus for the clinical utility of STREAM, for they can be quickly and accurately estimated by therapists without need for measuring each range by goniometer. In addition, the broad middle range that includes any movement that is not complete, solves the problem of how to deal with patients who have active movement, but whose movements are influenced by tone and, therefore, have somewhat variable amplitudes. The movements of these individuals would generally fall in the mid-range, greatly simplifying the measurement in terms of amplitude.

Similarly, the quality of the movement is assessed simply as normal (or near normal) or abnormal. In addition, information from categories **'1a'** (partial amplitude; abnormal pattern), **'1b'** (partial amplitude; normal pattern) and **'1c'** (full amplitude; abnormal pattern) may provide insight into the cause of the movement deficit, with category **'1a'** potentially reflecting combinations of paresis and tone/abnormal control, category **'1b'** potentially reflecting mainly paresis, and category **'1c'** potentially reflecting tone/movement control problems. These potential associations will need to be studied further. If STREAM does differentiate between the various background causes of movement disorders observed following stroke, it will make an important contribution to the evaluation and treatment of hemiplegia.

Finally, although both physical independence and safety in performing functional activities are important for functional purposes, in some cases variables such as cognition and perception may play a greater role than motor recovery in a patient's functional ability. Therefore, it was decided that on STREAM, if a patient had the motor ability to achieve a given task, they would be given full credit for their performance even if supervision was required due to cognitive or perceptual deficits. To test a patient's functional independence, where the contributions of motor, cognitive

and perceptual functioning are considered simultaneously, other more suitable measures should be selected.

In summary, the scoring of STREAM was designed to reflect both the gross quality and the amplitude of active movement, and, for the mobility activities, physical independence as well. Moreover, the scoring was designed to be simple, unambiguous, and objective so that therapists would be able to carry out the testing accurately and easily, with no special training beyond having read the scoring instructions included in the test manual. To achieve these qualities, we have carefully selected and defined (see glossary of scoring terms, Appendix 8.3) the terms used in the scoring scheme.

5.1.4 The STREAM scoring form and test manual

The completed STREAM is comprised of thirty items, equally distributed over the three subscales. Equal weight is given to each subscale, and therefore to each item. To facilitate test administration, the items are listed on the scoring form in a convenient order moving from items tested in lying, to sitting, and standing. In addition, for quick reference, a table summarizing the scoring directives is included on the front of the STREAM scoring form. Movements are performed on both sides so that the therapist can compare and contrast both the quality and amplitude of movement on the unaffected side with that of the affected limb, and in this way judge whether the movement is grossly normal for a given individual, and whether the patient has understood the instructions. Also, the standardized verbal instructions to patients, and specific notes for each item are given directly on the clinical scoring form. A comprehensive instruction manual has been developed (Appendix 8) to further standardize the testing procedures, and to provide information on the conceptual basis and reliability of the instrument. The test manual also includes a subscale scoring form (Appendix 8.3) to facilitate obtaining scores for the individual subscales and for the total STREAM. This form has spaces to accumulate item scores under the appropriate subscale, and guides the user through transforming subscale scores to be out of 100, and deriving total STREAM scores (out of 100) by summing the transformed subscale scores and dividing by three.

Although the STREAM is relatively comprehensive, it is also quick and easy to administer, requiring only about ten to fifteen minutes of a therapist's time, and no special equipment or training. These features greatly enhance the STREAM's utility, particularly for use in the clinical setting, where standardized outcome measures are infrequently used because they are lengthy and/or cumbersome to employ.

5.2 Internal Consistency

The completed STREAM and its subscales all demonstrated excellent internal consistency, with Cronbach's alphas above 0.96 (extremity subscales both 0.979; mobility 0.965; total STREAM 0.984). Although the alpha coefficients for the thirty item instrument were, with the exception of the alpha for the lower extremity subscale, consistently slightly lower than the alphas obtained on the forty-three item instrument, this was not unexpected as alpha is influenced by the number of items included.

In addition, seven of the forty-three items on the test version had item to subscale correlations below 0.80 (and correspondingly low correlations with the total STREAM scores), indicating a relatively weak association with the subscale to which they belonged. The five items with the lowest item to subscale correlation were eliminated during the process of item reduction; the two items that were retained despite relatively low item to subscale correlations (ie. 0.61 and 0.59) were basic mobility items--specifically, rolling and bridging. These two items were retained because they met all our criteria for item retention; they were measuring unique movements that the panels had considered important aspects of motor function, and they were moderately correlated with the other test items and with the total scale (0.64 and 0.58) so were related but not redundant. The alpha coefficients for the mobility subscale increased only slightly, from 0.965 to 0.969, when calculated with each of these items removed. Thus, the retention of rolling and bridging only minimally detracted from the overall internal consistency, but is reflected in the slightly lower alpha obtained for the mobility subscale. In contrast, the slight increase in the alpha for the lower extremity subscale, from 0.972 on the test version to 0.979 on the thirty item STREAM, may be explained by the fact that the two items that were eliminated from this subscale had low correlations (0.50 and 0.71) with the subscale total; because they did not have a strong association with the subscale, removal of these items improved the alpha for the subscale.

The alpha coefficients for STREAM and its subscales surpass the recommended 0.90 (Feinstein, 1987) for an instrument to be clinically useful for measuring a specific concept, thus supporting the use of STREAM for evaluating the movements of stroke patients. The subjects included in the internal consistency analysis came from in-patient rehabilitation as well as acute and LTC settings, and their scores on STREAM were distributed across the entire range of possible scores. Therefore, the results of this analysis should be representative of the performance of STREAM when used with

the population for which it is intended. The high degree of internal consistency indicates that the items included in STREAM are measuring one concept, presumably the recovery of motor function. The STREAM, thus, is internally consistent. In addition, the high level of internal consistency adds further support to the appropriateness of the content.

5.3 Reliability Studies

The reliability of a clinical scale is essential, as this property is a prerequisite to both the validity and responsiveness of the instrument. Reliability is important so that an instrument can accurately reflect the concept being measured, and so that it can detect small clinical changes. Reliability coefficients of 0.95 or better are recommended as the minimal requirement for a clinical outcome measure used in decision making about individuals (Helmstadter, 1964; Nunally, 1978; Weiner and Stewart, 1984). The estimates of reliability obtained for STREAM under the conditions imposed in this study exceeded the required level, and support further testing of the measurement properties. The extent to which the results of these studies can be generalized to clinical practice is related to the characteristics of the subjects and raters involved, and to how closely the approach used to evaluate reliability approximates realistic testing conditions. In this section, the generalizability of the findings of the two reliability studies are discussed in relation to the populations involved, and some of the interesting findings that may impact on clinical practice or research are highlighted. Next, the differences in the results of the two reliability studies are compared to one another. Subsequent discussion compares the testing done on other related measures, and considers the generalizability of the results of these studies in relation to the various approaches that are used for evaluating reliability.

5.3.1 Direct observation reliability study

In this study, pairs of raters from a group of six participating therapists simultaneously and independently used STREAM (ie. one rater observing the other doing the assessment) to evaluate the movements of twenty individuals who were undergoing active rehabilitation for stroke.

The clinical profiles of the twenty subjects participating in this study (summarized in Table 4.7) were diverse, and reflected the distribution and range of comorbid medical problems typically encountered in the stroke rehabilitation setting. The subjects demonstrated a wide range of motor ability, with total scores on STREAM ranging

from 23/100 to 100/100, so the findings of this study should be generalizable across these scoring ranges for similar in-patient rehabilitation populations. Further study is needed, however, to determine the reliability of STREAM in other stroke populations and clinical settings, such as acute or LTC.

The clinical backgrounds of the six raters were diverse: they had graduated from three different universities, worked at three different facilities, and had a range of from two to nine years experience as therapists, and several (1.5 to 3.5) years of experience working in stroke rehabilitation. Very limited training was given on the use of the STREAM; raters attended a short session during which the methods to be followed during the study were explained and a videotaped STREAM assessment was viewed and briefly discussed; they were also asked to do practice evaluations on two patients using STREAM. Thus, similar levels of reliability should be achievable in clinical practice simply by having therapists read the test manual and administer the STREAM a few times to become familiar with the instrument.

For our sample of twenty subjects, there was considerable variability in the scores on all the retained items. The distributions of scores on individual items are shown n Appendix 9.1. Only one score category, '**1b**', received very few ratings over all items; that is, patients who were only able to perform a movement partially, were also likely to use abnormal movement patterns. Other interesting but not surprising observations were made regarding the distribution of agreement for the ratings given within category '**1**' (shown in Table 4.9). Specifically, disagreements within category '**1**' scores occurred between '**1a**' and '**1b**' (qualitative) or '**1a**' and '**1c**' (quantitative) twice as frequently as between '**1b**' and '**1c**' (qualitative and quantitative). Also, disagreements by two categories occurred only within the mobility subscale, and were exclusively between categories '**1**' and '**3**'. Five of the six disagreements between these two categories involved '**1c**' (complete and independent movement, but abnormal pattern), reflecting the therapists' relative difficulty in differentiating between normal and abnormal movement patterns for some movements. The remaining disagreement was due to a rating of '**1b**' (partial assistance required, but normal pattern) given for bridging, probably reflecting a problem in deciding if partial assistance was required for stabilizing this patient's knees in mid-line for this movement. In these cases where the discrepancy was of two categories, the differences between the scores related only to one aspect of movement--either the quality of the movement, or the quantity. Obviously, these particular two category discrepancies are less serious than would be

the case, for example, if the difference in scores related to both quantity and quality of the movement (eg between **'1a'** and **'3'** or **'0'** and **'2'**), and fortunately, this never occurred.

Only one of the items retained on the completed STREAM demonstrated less than excellent reliability; kappa was only 0.65 for the item measuring sit to stand. This item was kept because the consensus panels had deemed it to be crucial as an important milestone of motor functioning, and because it performed well in terms of internal consistency, correlating at 0.83 with the subscale score and 0.77 with the total score. The disagreements between raters ratings on this item (on both reliability studies) were almost exclusively between categories **'1c'** and **'2' or '3'**. That is, the raters had difficulty differentiating between normal and abnormal movement patterns on this item. The instruction for this item was to "*stand up; try to take equal weight on both legs*"; the intention of the cue was to encourage equal weightbearing thus eliciting the patient's best performance. It is however difficult for the raters to discern the actual distribution of weight during the movement. The reliability of this item may be improved by adding a note to raters explaining the intent of the cue to patients, and asking them to grade the movement by observing the symmetry in both the vertical and horizontal planes. For example, "**Note:** asymmetry such as trunk lean, trendelenburg, hip retraction, or excessive flexion or extension of the affected knee = marked deviation (score **1a** or **1c**)".

Of the 1,200 ratings given in the direct observation reliability study (20 subjects x 2 raters x 30 items), forty-seven ratings of **'X'** were given. One patient was unable to perform eleven items (two lower extremity, and nine mobility items) due to pain on weightbearing, and dizziness in standing; for each of the remaining six subjects with missing data, only a few items (four or less) could not be performed hence were given ratings of **'X'**. Reliability was not adversely influenced by the missing data; excluding these individuals would not have further improved reliability. The effect of the missing data on the validity and interpretation of subscale and total STREAM scores, however, is still to be determined. A maximum needs to be set, through studies of validity and responsiveness, for the number of missing item scores allowed for subscale and total STREAM scores to still be meaningful.

Of the three subscales, the upper extremity subscale was the most reliably scored, followed by the lower extremity and basic mobility subscales (GCCs were 0.994, 0.993, 0.982). The finding of slightly greater reliability for the upper extremity

subscale, may be due to a greater heterogeneity in patients' upper extremity scores, as upper extremity recovery tends to be slower and less complete than that for the lower extremity. Another possible contributor to this slightly higher reliability may be that several of the patients had flaccid upper extremities, and clearly were not able to perform the test movement at all (so received a score of '**0**') thereby reducing the possibility of rater disagreement on scoring.

Over the course of this study, it was noted that several subjects did not conform to the expected pattern of greater improvement in the leg and mobility than in arm function. Interestingly, patients who were still in hospital many months following their stroke, and who were primarily wheelchair dependent, tended to demonstrate considerable recovery in arm function relative to the leg and mobility subscale scores. A recent study (Duncan et al 1994), involving a population similar to the subjects in the STREAM study, compared the rates of recovery from motor impairments for the upper and lower extremities (as measured by the Fugl-Meyer limb subscales). In line with the findings of the STREAM study, their findings show that one group of stroke patients, those with moderately severe initial motor impairments, tended to recover to a greater extent in the upper extremity than in the lower extremity. The relationships between the recovery from motor impairments and recovery of function in the limb and mobility domains, the severity of the initial impairments, and the duration of time since stroke, warrants further investigation.

Another interesting but unexpected finding made during the study was that, where the observing raters' scores differed from the scores given by the rater doing the actual hands on assessment, they generally did so by giving a higher score. Although this difference in scores was not significant, it was consistent, and is apparent from the plots of scores for this study, shown in Figure 4.4. There is no obvious explanation for this incidental observation of higher scores given by the observers. In clinical practice, however, this will not be an issue, as therapists will do the actual hands on assessing when scoring STREAM.

And finally, another interesting and promising finding in this study was that the GCCs were identical for the raters fixed and raters as random facets models, indicating that the raters contributed little variability to the scoring of STREAM. This means that, theoretically, reliable STREAM scores should be attainable by the general population of physical therapists, and not just by the selection of raters in this study.

5.3.2 Videotaped assessments reliability study

In this study, twenty therapists viewed and rated four videotaped STREAM assessments on two occasions, and estimates were obtained for intra-rater as well as inter-rater reliability (on each of the two occasions). The clinical profiles of the four videotaped subjects participating in this study were relatively diverse, and reflected the distribution and range of comorbid medical problems typically encountered in the stroke rehabilitation setting. These subjects demonstrated a wide range of motor ability, with scores on STREAM ranging from 21/70 to 65/70, thus, the findings should be generalizable across these scoring ranges. The twenty therapists involved in rating the tapes had graduated from six universities, were working in twelve different facilities, and had a wide range of general (1 to 33 years) and stroke specific (0.5 to 11 years) clinical experience. The diversity of the clinical backgrounds of the participating therapists supports the generalizability of the results of our study across the spectrum of training and experience. In addition, the raters received very little training in the use of the STREAM prior to the videotape viewings. Two weeks prior to the study they were given a copy of the STREAM and scoring instructions, and were asked to do two practice evaluations of patients using STREAM. Despite this very limited training, excellent reliability was achieved, suggesting that similar levels of reliability should be achievable in the clinical setting with no formal training save for reading the test manual. The limitations of the results of this study due to the use of videotaped assessments are presented below in the sections comparing the various approaches for testing reliability. The remainder of this section discusses some of the interesting findings of this study.

The distribution of agreement on this study was similar to that of the direct observation reliability study. That is, the disagreements of two categories again occurred almost exclusively within the mobility subscale, and were primarily between categories '**1**' (mainly '**1c**') and '**3**'. In contrast to the previous study examining inter-rater agreement, where one therapist's idea of normal may differ from another therapist's definition of normal, in this study the same therapist gave these different ratings on two occasions, illustrating that agreement on what is normal or abnormal movement was occasionally a problem not only between therapists, but also for individuals on different occasions.

We ranked each of the tapes based on their levels of perfect agreement (see Table 4.15). Not surprisingly, slightly lower levels of overall agreement were observed for

tapes C and D, two subjects that were moderately influenced by tone and that had perceptual problems. This finding suggests that tone and perceptual problems may make the scoring of movement slightly more difficult. Our highest functioning individual was most reliably scored on STREAM, as this individual could clearly perform the majority of the movements well, so deciding on the score to give was not a problem for therapists. The ranks of agreement on each of the tapes differed slightly for the upper extremity subscale. That is, the patient who had the highest rank for level of perfect agreement had a flaccid arm, thus was generally more easily scored for the upper extremity, as it was clear that category **'0'** applied.

Within category **'1'**, disagreements between categories **'1a'** and **'1b'** (qualitative) and categories **'1a'** and **'1c'** (quantitative) occurred more frequently than between **'1b'** and **'1c'** (qualitative and quantitative) for the limb movements. In contrast with the direct observation study, for mobility activities there was an equal distribution of disagreements. That is, on average over the four tapes, raters were equally likely to disagree in terms of the movement quality, quantity, or both at once. This result may be in part due to the use of videotapes, from which it may be harder to observe movements both in terms of quality and quantity. The subject selection may also have had an influence on the distribution of disagreements observed within category **'1'**. The discrepant ratings were generally isolated to a few specific items. The disagreements between **'1b'** and **'1c'** occurred almost exclusively on the tape A upper extremity subscale (of the seven discrepant ratings, two were for supination/pronation, and three were for closing the hand), and on the tapes B and C lower extremity subscales (of nine discrepant ratings, all were given on the leg movements performed in sitting). Either these particular items were not clearly shown on these videotapes, or the patients performances of these items were borderline in terms of which score they should be given, and may have been as difficult to rate consistently while directly observing the patients performance in the clinical setting.

It is obviously more difficult to view clearly smaller movements on videotape, such as movements of the hand, than to view the larger range movements of the extremities and basic mobility subscales. Therefore, because of the greater number of movements included in the upper extremity subscale that are potentially difficult to view, the finding of slightly lower reliability for this subscale on the videotape study is not surprising.

Interestingly, there was a tendency for raters to score the videotaped subjects higher on the second ratings, and in fact for two of the tapes this difference was significant (signed ranks; $p<.05$). It is possible that this trend was an artifact of the therapist's 'learning curve', and that, had a third viewing session been carried out and compared with the scoring of the second session, this trend would no longer be evident. It would be of interest to examine what effect further training would have achieved on the variability of the scoring.

5.3.3 Comparison between the two reliability studies

The use of two different methods to evaluate reliability (direct observation and videotape viewing) generated a considerable amount of information. It also provides an opportunity to contrast these two approaches and examine some of the differences between these two methods for testing reliability. The two studies differed in general design, in that the direct observation study was aimed at estimating the agreement within raters, while the videotaped study was aimed at estimating the agreement between raters (although the agreement within raters was also evaluated for each of the two occasions). Thus, a comparison of these two studies is a bit like comparing apples to oranges. None the less, a number of interesting differences between these two approaches to measuring reliability were evident. The distributions of agreement, signed rank statistics, and GCCs for subscales and total scores for each of the two studies are summarized side by side for comparison in Table 5.1.

In general, the results of the two reliability studies were comparable, with both studies indicating excellent overall agreement on the scoring of STREAM. The most notable difference between the two studies was the slightly lower generalizability coefficient for the upper extremity subscale in the videotaped assessments study, which likely can be explained at least in part by the different methods used. Also, not surprisingly in light of the more controlled testing conditions, overall inter-rater agreement on the videotaped assessments reliability study was slightly higher than on the direct observation study. The finding of slightly higher intra-rater than inter-rater agreement is also as would be expected, as typically agreement within raters is better than that between raters. The comparison of these two studies would indicate that, although the results of the two reliability studies were comparable, the study methods do influence the results, and therefore must be considered when judging the reliability of individual test items, subscales, and of an instrument as a whole.

TABLE 5.1
Comparison of the Results of the Two Reliability Studies

	DIRECT OBSERVATION	VIDEOTAPED ASSESSMENTS				
		Tape: A	B	C	D	Combined
SUBSCALES						
Perfect Agreement						
Upper Extremity	88.3%					88.4%
Lower Extremity	91.2%					86.5%
Basic Mobility	88.7%					86.4%
Signed Ranks[a]						
Upper Extremity	NS	NS	NS	S	S	
Lower Extremity	NS	NS	NS	S	S	
Basic Mobility	NS	NS	NS	S	S	
GCCs						
Upper Extremity	.994					.963
Lower Extremity	.993					.999
Basic Mobility	.982					.999
TOTAL SCORES						
Perfect Agreement	89.4%					85.7%
Signed Ranks	NS	NS	NS	S	S	
GCCs	.995					.999[b]

a : 'NS' indicates good agreement (ie.non-significant signed rank, $p > .05$)
'S' indicates poor agreement (ie. significant signed ranks, $p < .05$)
b : the same value was obtained for inter- and intra-rater agreement

5.3.4 Overview of reliability studies done on measures of motor recovery

Table 5.2 provides a comprehensive summary of the types of reliability studies that have been published to date on the available stroke motor assessments. More details are given in the literature review, in section 2.2.3, for the reliability studies carried out on the Fugl-Meyer Assessment, the Motor Assessment Scale (MAS), and the Chedoke-McMaster Stroke Assessment (C-McSA). The STREAM, although published only in abstract form as yet, is included in the table to facilitate a comparison of the extensiveness, the quality, and the results of the STREAM reliability studies, with studies done for related measures. As can be seen, the studies varied considerably in terms of: the types of reliability (inter-rater, intra-rater, test-retest, and/or internal consistency) evaluated, the general approach used (direct observation or videotapes), the numbers of patients and raters involved, the types of patients assessed

(rehabilitation or chronic), and the type(s) of analysis conducted. Within the direct observation type studies, there were also differences in how the patients were rated, such as whether each of the raters performed the assessments on separate occasions (with a range of intervals between ratings), or both raters were present for the same performance, with one (or both) rater(s) observing the assessment.

In brief, only two of the measures (including STREAM) were evaluated for internal consistency as well as for rater agreement, but most of the published instruments have been evaluated in terms of both inter- and intra-rater agreement. Only two instruments (the Fugl-Meyer and MAS) have been evaluated twice (using different populations, methods and analyses) for inter-rater agreement.

The majority of published studies (ten of the eleven reported here) used direct observation in the clinical setting to assess reliability, four studies used videotaped assessments in addition to direct observation (for determining intra-rater agreement in two studies, and for inter-rater agreement in the remaining two), and one study used only videotaped assessments (and determined intra-rater agreement only). Videotaped assessments were used to ensure the stability of the performance being rated over time (for evaluating intra-rater reliability), and/or to minimize burden to patients for repeated testing. It was also generally recognized that these videotaped reliability studies were preliminary in nature, and represented reliability under conditions that were less variable that in typical clinical practice. Of the eight direct observation type studies for evaluating inter-rater reliability, four involved raters evaluating subjects on separate occasions (with intervals ranging from three hours to fourteen days between occasions), and four involved raters assessing in pairs, with one (or both) rater(s) observing the assessment. In two of the four 'observer' studies, one rater observed the evaluation being done by the other rater, while in the remaining two studies both raters observed while another therapist performed the evaluation. In light of the findings of the STREAM direct observation reliability study, where observing raters tended to give higher scores than the rater performing the assessment, and because it is unusual in the clinical setting for therapists not to be doing their own hands-on evaluation while rating a patient, this way of testing may have questionable relevance. Having the therapists perform separate evaluations may more directly reflect how these assessments are generally used in the clinical setting, with different raters carrying out evaluations at different times, and multiple possible sources of variability. There are, however, several important reasons for having one rater observe while the other therapist

performs the assessment, including: convenience, to minimize the burden to patients, and to ensure that an identical performance is being rated by both raters. By eliminating the possibility of variability of the patient's performance from the equation, a relatively 'pure' estimate is obtained for the agreement between raters on the scoring of the instrument. 'Pros' and 'cons' exist for each of the above methods, and, unfortunately, some degree of compromise is usually necessary. Future studies could feasibly do two studies--one with raters carrying out assessments independently on separate occasions, and another with raters simultaneously rating the same performance. Not only would this provide additional evidence for reliability, it would also allow for a comparison to be made between these two methods, as was done in this study for the direct observation and videotaped approaches.

Only three studies looked at intra-rater reliability in the clinical setting using direct observation rather than videotapes. However, two of these studies (Duncan et al, 1983, and Lincoln and Leadbitter, 1979) included chronic stroke patients only, looked at intra-rater agreement for only one rater, and used Pearson's correlations (reflecting only a linear relationship) for the analysis rather than the more informative ICC (reflecting actual concordance of scores). The other direct observation intra-rater study (Corriveau et al, 1992) involved rehabilitation patients, in an attempt to obtain estimates more generalizable to the rehabilitation setting. In this study, nineteen therapists carried out assessments on three occasions within a two week interval. Rehabilitation patients may be expected to improve during this interval, and therefore scores would potentially change over the course of the study. It is also possible that therapists would remember the scores given on the previous occasion. Although estimates of intra-rater reliability in the rehabilitation setting are required for instruments used in this setting, obtaining these estimates by the direct observation method requires several compromises that may bias the results, and so the videotaped assessments method, although artificial, may be a reasonable alternative.

TABLE 5.2

Summary of Reliability Studies Done for Stroke Motor Assessments

Study	Methods	Subjects	Raters	Results
STREAM	**Inter-rater, and internal consistency**			
This study	Direct observation (1 rater observing)	20 rehab patients	2 (of 6)	GCC=.99 (u/e .99; l/e .99; mobility .98) Chronbach's alpha=.98 (u/e .98; l/e .98 mobility .97)
	Intra-rater			
	Videotapes (2 occasions; 1 month interval)	4 rehab patients	20	GCC=.99 (u/e .96; l/e .99; mobility .99)
C-McSA	**Inter-rater**			
Gowland, 1993	Direct observation (1rater observing) *in first week of admission to rehab (videotaped)	32 rehab patients	2 (of 6)	ICC=.99
	Intra-rater			
	Videotapes (2 occasions; >2 week interval) *impairment inventory only	32 rehab patients	3	ICC=.98
	Test-retest			
	Direct observation (1rater observing) (2 occasions; >5 day interval) *disability inventory only	32 rehab patients	2 (of 6)	ICC=.98

Note: First author's name given only.

TABLE 5.2 (cont)

Summary of Reliability Studies Done for Stroke Motor Assessments

Study	Methods	Subjects	Raters	Results
Fugl-Meyer	**Inter-rater**			
Sanford, 1993	Direct observation (assessed separately; same or next day)	12 rehab patients	3	ICC=.96 (u/e .97; l/e .99)
Duncan, 1983	**Inter-rater**			
	Direct observation (assessed separately; same day) *evaluating extremity subscales only	8 patients on u/e; 10 patients on l/e subscale	3 (of 4)	Pearson's r= u/e .99; l/e.89
	Intra-rater			
	Direct observation (3 occasions; 3 week intervals)	19 chronic CVA (>1 year post CVA)	1(of 1)	.99 for total score (.86-.99; subscales)
MAS	**Inter-rater**			
Poole, 1988	Direct observation (raters both observing)	24 rehab patients	2	Spearman's r_s=.99
Loewen, 1988	**Intra-rater**			
	Videotapes (2 occasions; 1 month interval) *excluded tone	7 rehab patients	14	Kappa 80%>.75 r_s=.83-1.00
Carr, 1985	**Test-retest**			
	Direct observation (2 occasions; 1mos interval)	14 rehab patients	1(of 1)	Pearson's r=.98
	Inter-rater			
	Videotapes *agreement with criterion score	5 rehab patients (5 mos-34 yr post CVA)	20	Pearson's r=.95

1 Note: First author's name given only.

TABLE 5.2 (cont)

Summary of Reliability Studies Done for Stroke Motor Assessments

Study	Methods	Subjects	Raters	Results
Bobath	**Intra-rater**			
Corriveau, 1992	Direct observation (3 occasions; >48hr interval; within 14 days)	19 rehab patients	1(of 19)	ICC=u/e .79; l/e .77
	Inter-rater			
	Direct observation (assessed seperately; >48hr interval; within 14 days)	18 rehab patients	3(of 3)	ICC=u/e .97; l/e .95
Rivermead	**Inter-rater**			
Lincoln, 1979	Videotapes *original version UE section revised in response to findings of this study	7 chronic CVA	7	ANOVA (sig for u/e)
	Intra-rater			
	Direct observation (2 occasions; 1 month interval)	10 chronic CVA	1(of 1)	Pearson's r=.66-.93
LaVigne	**Inter-rater, and internal consistency**			
Brosseau, 1992	Direct observation (same day; >3hr interval)	36 acute CVA	2 (of 4)	ICC=.77 u/e; .65 l/e; hand .87 Chronbach's alpha: u/e .91; hand .94; l/e .86
Physical Assessment for Stroke				
Ashburn, 1982	**Inter-rater**			
	Direct observation (paired observers)	15 rehab patients	2 (of ?)	No statistical analysis; disagreements on items noted.

1 Note: First author's name given only.

5.4 General Comparison of STREAM with Other Measures of Motor Recovery

The following section provides a general summary of how the known and expected characteristics of STREAM 'measure up' to related instruments. An overview of the characteristics of the published instruments for evaluating motor function after stroke was given in Table 2.1. In this table, the content domains, conceptual framework, scoring, administration time, and extent of evaluation of the psychometric properties of each measure were noted. Below, each of these characteristics are briefly summarized for STREAM, and compared with those of the other instruments.

5.4.1 Content domains

Two content domains are included in the STREAM: voluntary movements of the limbs, and basic mobility. Most of the instruments related to STREAM measure similar aspects of movement, although several instruments include additional domains, such as: sensation, pain, range of passive movement, and/or tone. While these additional domains are necessary for treatment planning purposes, they do not specifically reflect motor recovery and, therefore, should not be included in an outcome measure for this trait.

5.4.2 Conceptual framework

Several of the stroke motor assessments were based on the synergic patterns of motor recovery described by Brunnstrom (1970). Similarly, one measure (Guarrna et al, 1988) is based on Bobath's phases of recovery. In contrast, the MAS has taken a purely functional approach, and conforms to motor control principles. The conceptual frameworks of the remaining instruments were not specified, although they appear to fall somewhere between the synergic and the functional approaches. STREAM also employs an eclectic conceptual basis. That is, although the synergies are not specifically addressed in the STREAM's approach to measuring movement, they are reflected by the movements included in STREAM in as much as movements range from very basic (typically in synergic pattern--eg. hip and knee flexion) to gradually more difficult (and out of synergy--eg. knee flexion with hip extended). The STREAM movements are also all closely related to functional movements. For example, flexing the knee in sitting is required in preparation for standing, hip flexion in sitting is required for donning shoes and socks and placing the foot on a foot pedal if in a wheelchair, and reaching to touch the top of the head is required for grooming. Although functionally oriented, the movements evaluated on STREAM are intentionally limited to simple movements rather than tasks involving the manipulation of props (eg.

shoes, hair brushes, etc). There were two reasons for confining the movements to those that did not require equipment: 1) to minimize the potential confounding influence of perceptual and cognitive functioning on the movement, and 2) to enhance the portability and utility of STREAM. Hopefully, the approach to measuring movement employed by the STREAM will satisfy the proponents of the synergies, as well as therapists leaning towards a more functional approach.

5.4.3 Scoring

The related stroke motor assessments all use ordinal scales for the scoring of individual test items. The scales used range from two to seven points, and thus vary in terms of complexity and preciseness. Section 5.1.2 of this discussion outlined several issues related to the scoring dimensions (such as quantity, quality, or independence of movement) incorporated into the related instruments, and how these schemes influenced the scoring scheme developed for STREAM. In brief, several problems exist in the scoring of the other instruments: all of the dimensions of movement of interest to therapists (ie. quantity, quality, and independence of movement) are not adequately reflected; dichotomous scaling does not provide adequate sensitivity, while seven-point scales (particularly where descriptions for each level on each item are different) are excessively complex; and/or scoring categories are not objectively defined. The scoring of STREAM is intended to resolve these problems. Amplitude, gross quality, and independence in mobility are incorporated. Simple three point (limb movements) and four point (mobility) ordinal scaling is used, with consistent and objective descriptions applied across all items. The findings of this study, including the exceptional reliability and minimal administration time, would indicate that the STREAM scoring scheme has adequately resolved most of the problems identified in the scoring of the other instruments, while still maintaining simplicity and objectivity.

5.4.4 Administration time

A common characteristic of all of the stroke motor assessments is that they are intended to be administered by therapists, and require the therapists to observe and score the patient's actual physical performance rather than relying on self-report. The STREAM, however, is administered in about half the time required by the other measures (ie. in approximately fifteen minutes). Only one other instrument, the MAS, is carried out in a similar amount of time, and this instrument consists of only nine items. Although the STREAM is comprised of thirty items, the simple scoring process and the way in which the instrument is organized (ie. the flow of items from supine to

standing and from low to high level in terms of motor ability, the standardized instructions, and the fact that no special equipment is required) combine to facilitate rapid assessment. This quality of STREAM will certainly be much appreciated by clinicians.

5.4.5 Reliability

The reliability, and the quality and extent of the tests for this property, for STREAM and related clinical measures of stroke motor function, were summarized in Table 5.2, and discussed in section 5.2.4. To date, the inter- and intra-rater reliability, as well as internal consistency of STREAM have been evaluated. Two studies were carried out, a direct observation and a videotaped assessments reliability study, using sound scientific methods, and appropriate statistical analysis. These studies have provided a considerable volume of information on the reliability of STREAM, and, although relatively preliminary, the excellent results shown in these two studies are extremely promising.

Surprisingly, only one of the other measures (the LaVigne Motor Assessment) was evaluated in terms of internal consistency. This characteristic is a desirable quality for an instrument that is comprised of a number of items and intended to measure a particular concept. Moreover, this aspect of reliability can conveniently be obtained using the same data used to obtain estimates of rater agreement, provided that the subject's scores span the range of the instrument. And, the information obtained through internal consistency analysis can be used to support the appropriateness of the content. The exceptionally high level of internal consistency of STREAM is not likely to be outshone by other measures. It is expected that a study of the components of STREAM will further substantiate the appropriateness of the content, although this remains to be confirmed.

Several factors probably contributed to the very high estimates of reliability obtained for STREAM. The potential effects of the somewhat controlled conditions have already been elucidated. However, several characteristics of the STREAM, most notably the simple scoring scheme and standardized testing instructions, are likely to enhance its reliability, regardless of the testing conditions. In addition, the fact that the reliability on STREAM was very good across the full scoring range, as shown by the plots of paired ratings (Figure 4.4 for inter-rater agreement, Figure 4.5 for intra- and inter-rater agreement), is a further positive attribute of STREAM. This is an important quality for instruments intended to be used for evaluating patients with a wide range of

motor capabilities, and it is not clear from the literature whether the other stroke motor assessments demonstrated reliability across the scoring range.

5.4.6 Validity

Only two other instruments, the Fugl-Meyer and Chedoke assessments, have been evaluated in terms of content, and these only in single relatively preliminary studies. Furthermore, the methods used in the development and refinement of the other measures of stroke motor function are only vaguely described in the literature (if described at all), and unlikely to have matched the structured processes, including the consensus panels and item reduction phases, that contributed to the refinement of STREAM. In light of the information available to date, the STREAM may well surpass other measures in terms of support for content validity.

Few of the published instruments have been evaluated to any extent in terms of criterion validity. The Fugl-Meyer is generally accepted as the standard to which other measures are compared, as it has the largest body of evidence supporting its validity. The Chedoke scale is also gaining acceptance as a valid measure of motor impairment and disability for stroke. The extent to which STREAM compares to these related measures, both in terms of current and predictive measurement, remains to be evaluated. However, because the structure of STREAM is similar to the 'silver standards' available (ie. the Fugl-Meyer, and Chedoke assessments), if these instruments are used as criterion measures in evaluating the reliability of STREAM, a relatively strong relationship is expected. Hopefully evidence related to this aspect of STREAM's validity will be available in the near future, and will be further supported by the relationship of STREAM scores with laboratory measures of motor function, or global functional measures (ie. construct validity).

Construct validity relates to the extent to which the recovery of movement is being measured by the instrument. The Fugl-Meyer scale is the only instrument for which this psychometric property has been extensively studied, although a few early studies of construct validity have been published on other measures. Evaluations of the relationships between STREAM scores and other assumed indicators of motor recovery need to be carried out to provide additional evidence about the STREAM's ability to reflect the concept of motor recovery following stroke.

5.4.7 Responsiveness

To date, no instrument has adequately met the needs of therapists at all of the levels of stroke rehabilitation--which may explain the continued quest and identified need for a measure that will be acceptable for use across the full range of motor impairments ensuing stroke. Responsiveness to clinical change in motor function is imperative for an instrument intended to be used as an outcome measure for motor recovery in stroke. Only the Fugl-Meyer and Chedoke assessments have been evaluated for responsiveness, and only to a limited extent: ceiling effects have been identified as a problem common to both measures, and their sensitivity to changes in the very low level patient have not been evaluated.

Although the responsiveness of STREAM has not yet been formally evaluated, several factors are likely to have favorably influenced this property. During the development phase of STREAM, panel participants were selected to represent all phases of stroke rehabilitation. These individuals were asked to take the probable responsiveness of individual items to clinically important changes into consideration in deciding the merit of each item. Thus, the feedback from therapists working with all ranges of motor functioning influenced the STREAM's development. As well, the variability of scores over the thirty items that make up the revised STREAM, with twenty of the items scored on a three point, and ten items scored on a four point scale (for a total of $[(20 \times 3)+(10 \times 4)]=100$ score options), will allow ample opportunity for meaningful clinical changes to be noted.

The distribution of scores on STREAM would indicate that STREAM reflects differences in motor status both within and between individuals, and therefore possesses some degree of sensitivity. Patients in earlier stages of recovery generally received lower STREAM scores than those evaluated at a later stage, and patients in acute care and LTC facilities showed a tendency to score lower than individuals in in- or out-patient rehabilitation programs. In addition, only those individuals who demonstrated very near normal motor function received a perfect score on the STREAM, allowing room for improvement to be noted even for very high functioning patient. Only subjects that were essentially vegetative were likely to receive a total STREAM score of zero, as some very low level items have been included (such as rolling, shrugging shoulders), and because scores of '1' are given for partial movements. Hence, the STREAM appears to span the spectrum of motor function well, with little evidence of ceiling or floor effects. Feedback from therapists using STREAM to evaluate their

patients over time also indicates that STREAM scores parallel the observed clinical changes. Further study of this important measurement property has been initiated.

5.4.8 Overview of the comparison between STREAM and related measures

In summary, in all of the characteristics that are important for an outcome measure for the recovery of movement following stroke, the STREAM rivals other related measures. That is, it provides a considerable amount of highly internally consistent and reliable information on the movement of individuals following stroke; it shows promise of being a valid and responsive measure of motor recovery; and, it is extremely straightforward to administer, requiring minimal time, and no special equipment or training, thus, its clinical utility is excellent. In light of the above features, in many ways the STREAM offers a considerable advantage over the related measures.

5.5 Limitations

5.5.1 Limitations of this study

There were several limitations inherent to this study. Most notably, this study represents only an initial evaluation of the reliability of STREAM. Before STREAM can be advocated for use as an outcome measure for research in stroke rehabilitation, further studies will be required to: 1) provide evidence in support of the construct and criterion validities, and responsiveness of the measure to motor recovery, 2) investigate the reliability of the instrument in alternate settings such as acute care, extended care, or outpatient rehabilitation programs, and 3) to further substantiate the excellence of the reliability determined in this study.

Another potential limitation relates to the subject selection procedure, which was one of convenience. Attempts were made, however, to select a broad and representative range of a) patients with respect to severity and time since stroke, and b) therapists with respect to training, experience, and background. The characteristics of subjects and therapists are documented, and one can see that we did indeed include a wide range of subjects and raters, that would typically be encountered in stroke rehabilitation settings. Because the JRH admission criteria results in the exclusion of very mildly and very severely impaired patients, however, our subjects may not be fully representative of all stroke patients. Also, we were only allowed to recruit patients who were not involved in another research study that was being carried out simultaneously at the JRH. Since the patients that we recruited had generally been excluded from the other study due to

comorbid conditions, such as cardiac, perceptual or communication problems, our sample likely had a higher proportion of complications and comorbid conditions, and a slightly lower average functional level than would have been the case had we not been restricted in patient selection. Intuitively, the effect of this selection bias might be a more conservative reliability estimate, assuming that extremely high or low level patients are scored more easily (and therefore more reliably). The fact that no 'mid-range' reduction in reliability was evident under the conditions of testing in this study (see Figures 4.4 and 4.5), however, would suggest that STREAM is reliably scored regardless of the functional level of the subject.

The testing procedures employed in the direct observation inter-rater reliability study, where both ratings are made in the same testing session, in a quiet environment, represent the ideal conditions for achieving reliable scoring, as this ensures that variability in patient performance is not a contributor to measurement error. This is in contrast to the real world situation, where therapists test patients separately, on different occasions and under more variable conditions. Similarly, using videotaped assessments to estimate intra-rater reliability can only approximate the clinical reliability of the instrument. Videotaped assessments may be performed in a more standardized fashion and under more controlled conditions than would likely be found on a day to day basis in a rehabilitation setting, where therapists are frequently interrupted or distracted during assessment sessions. Reliability estimates obtained under these somewhat artificial conditions may represent 'optimal' reproducibility rather than the 'real world' situation, and provide only a preliminary estimate, or an estimate of maximal intra-rater reliability. Another drawback of using videotaped assessments is that smaller movements, such as movements of the hand, may be difficult to view clearly. And, it may be difficult to rate how much assistance a patient is actually being given. The generalizability of the results of this study to more realistic situations needs to be determined.

Finally, our sample sizes were relatively small, however, they should be adequate for preliminary study of the measurement properties. For the direct observation reliability study, if agreement had not been achieved over twenty subjects, increasing the sample size would not have further improved the agreement obtained. Cicchetti (1976) suggests a minimal sample size requirement, when using weighted Kappa in assessing the reliability of rating scales, of at least two times the square of the number of ordinal

categories. That is, for three scale categories, a minimum of eighteen subjects are required.

5.5.2 Limitations of the STREAM as a measure of motor recovery

The STREAM measures only one attribute of stroke (ie. motor recovery). For treatment planning, STREAM should be used with measures of other important attributes, such as spasticity, ROM, sensation, balance, and functional independence in ADL. For example, instruments such as the Ashworth Scale of Spasticity (Ashworth, 1964), the Berg Balance (Berg et al, 1989) and the Barthel Index (Mahoney and Barthel, 1965) could be used in conjunction with the STREAM to obtain a detailed profile of a patient's physical function.

Although the individual subscales on STREAM have demonstrated excellent internal consistency and reliability, the STREAM was designed to be used as a global or summary measure for the recovery of movement and mobility. Each subscale consists of only ten items and, until the validities of the individual subscales of STREAM are determined, the use of the subscales independently from the test as a whole cannot be advocated.

5.6 Implications for Future Research on STREAM

The STREAM is now ready for clinical use, and for further testing of its psychometric properties to support its usefulness as a clinical research tool. The involvement of stroke patients with varied clinical profiles in the reliability studies, has indicated that the instrument's reliability across a relatively diverse population is excellent. The reliability in acute care, long term care, and out-patient settings, however, remains to be tested. Further tests of reliability should be carried out in different clinical facilities and settings, and under less controlled conditions than was the case in this study, in order to show the generalizability of the STREAM's reliability across institutions, patient populations, and testing conditions. Further measurement studies are also required to evaluate criterion and construct validities, and longitudinal studies are needed to assess the responsiveness of STREAM to change. And, ultimately, the STREAM will need to be used in clinical trials and evaluated in terms of its efficiency relative to other instruments for discerning the effect of treatments on the recovery of movement.

5.7 Conclusions

Through the collaborative effort of our panels of physiotherapists, the STREAM has undergone a thorough refinement of its content. Close attention has been paid to the appropriateness of the items included, the scoring format, overall comprehensiveness for measuring motor recovery, and clinical utility. The revised STREAM has demonstrated excellent internal consistency, inter-rater, and intra-rater agreement, and is now ready for preliminary use in clinical settings. It is anticipated that the simplicity and overall clinical utility of the STREAM will facilitate the incorporation of this instrument into the clinical setting for the routine objective measurement of motor function. The findings of this study are encouraging, and suggest that STREAM warrants further testing of its measurement properties to determine its usefulness as an outcome measure. Contingent on similarly favorable results from further studies of the measurement properties, STREAM may become a key measure of the recovery of voluntary movement and basic mobility and thereby contribute to our understanding of the evolution of motor recovery following stroke.

REFERENCES

Anderson T, Baldridge M, Ettinger M. Quality of care for completed stroke without rehabilitation: evaluation by assessing patient outcomes. Arch Phys Med Rehabil. 60: 103-107, 1979.

Andrews K, Brocklehurst J, Richards B, Laycock P. The rate of recovery from stroke--and its measurement. Int Rehab Med. 3: 155-161, 1981.

Arsenault B, Cleather J. A perspective on Physiotherapy Canada; results of 1981 readership survey. Physiother Can. 34(1): 28-36, 1982.

Ashburn A. A physical assessment for stroke patients. Physiotherapy. 68: 109-113, 1982.

Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. Practitioner. 162: 540-542, 1964.

Badke M, Duncan P. Patterns of rapid motor responses during postural adjustments when standing in healthy subjects and hemiplegic patients. Phys Ther. 63(1): 13-20, 1983.

Berg K, Wood-Dauphinee S, Williams I, Gayton D. Measuring balance in the elderly: preliminary development of an instrument. Physiother Can. 41: 304-311, 1989.

Bergner M. Health status measures: an overview and guide for selection. Ann Rev Public Health. 8: 191-210, 1987.

Bernspang B, Asplund K, Eriksson S, Fugl-Meyer A. Motor and perceptual impairments in acute stroke patients: effects on self-care activity. Stroke. 18(6): 1081-1086, 1987.

Bisch L, Semenciw R, Wilkins K. Temporal and spacial patterns in stroke mortality and morbidity among the Canadian elderly. Chronic Diseases in Canada. 10: 63-67, 1989.

Bobath B. Adult Hemiplegia: Evaluation and Treatment. London: William Heinemann Medical Books Ltd., 1978.

Bogousslavsky J, VanMelle G, Regli F. The Lausanne Stroke Registry: analysis of 1,000 consecutive patients with first stroke. Stroke. 19: 1083-1092, 1988.

Bohannon R. Muscle strength changes in hemiparetic stroke patients during inpatient rehabilitation. J Neurol Rehab. 2: 163-166, 1988a.

Bohannon R. Is the measurement of strength appropriate in patients with brain lesions? Phys Ther. 69: 225-236, 1989.

Bonita R, Beaglehole R. Recovery of motor function after stroke. Stroke. 19: 1497-1500, 1988.

Bonita R. Epidemiology of stroke. The Lancet. 339: 342-344, 1992.

Borg W, Gall M. Educational Research: an Introduction edition 5. New York: Longman Inc., 1989.

Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the Intraclass Correlation Coefficient: towards the integration of two traditions. J Clin Epidemiol. 44: 381-390, 1991.

Broderick J, Phillips S, Whisnant J. Incidence rates of stroke in the eighties: the end of the decline in stroke. Stroke. 20: 577-582, 1989.

Brooks S. The neural basis of motor control. New York: Oxford University Press, 1986.

Brosseau L, Potvin L. Inter-rater reliability and construct validity of the LaVigne Motor Recovery Assessment Scale following a cerebovascular accident. Physiotherapy Theory and Practice. 9: 91-102, 1993.

Brunnstrom, S. Movement Therapy in Hemiplegia. New York: Harper and Row, 1970.

Carey J, Burghardt T. Movement dysfunction following central nervous system lesions; a problem of neurologic or muscular impairment? Phys Ther. 73: 538-547, 1993.

Carr J, Shepherd R, Nordholm L, Lyme D. Investigation of a new motor assessment scale for stroke patients. Phys Ther. 65: 175-180, 1985.

Carr J, Shepherd R. A Motor Relearning Program for Stroke. Rockville, Maryland: Aspen Publications, 1986.

Carr J, Shepherd R. Movement Science: foundations for physical therapy in rehabilitation. Rockville, Maryland: Aspen Publications, 1987.

Carroll D. A quantitative test of upper extremity function. J Chron Dis. 18: 479-491, 1965.

Cicchetti D. Assessing inter-rater reliability for rating scales: resolving some basic issues. Brit J Psychiat. 129: 452-456, 1976.

Clarke B, Gowland C, Brandstater M. A re-evaluation of the Brunnstrom assessment of motor recovery of the lower limb. Physiother Can. 35: 207-211, 1983.

Cohen J. A coefficient of agreement for nominal scales. Ed Psych Msmt. 20: 37-46, 1960.

Cohen J. Weighted Kappa: nominal scale agreement with provision for scale disagreement or partial credit. Psychological Bulletin. 70: 213-220, 1968.

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

Corcos D. Strategies underlying the control of disordered movement. Phys Ther. 71: 25-38, 1991.

Corriveau H, Arsenault A, Dutil E. An evaluation of the hemiplegic patient based on the Bobath approach; a reliability study. Disabil Rehabil. 14(2): 81-84, 1992.

Criak R. Abnormalities of motor behavior. In: Contemporary Management of Motor Control Problems: proceedings of the II Step Conference. Lister M.ed. Fredericksburg, VA: Bookcrafters, 1991. pp. 157-163.

Craik R, Oatis C. Gait assessment in the clinic: issues and approaches. In: Measurement in Physical Therapy. Rothstein J.ed. New York: Churchill Livingstone, 1985. pp. 169-201.

Crick T, Brennan R. GENOVA: a generalized analysis of variance system, version 2. Iowa, The American College Testing Program, 1983.

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

Cronbach L, Glester G, Nanda H. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley, 1972.

Dalkey N. Delphi. Rand Corporation, 1967.

Del Greco L, Walop W, McCarthy R. Questionnaire development: validity and reliability. CMAJ. 136: 699-700, 1987.

Delbecq A, Van de Ven A, Gustafson D. Group Techniques for Program Planning: a guide to nominal group and delphi processes. Dallas, Texas: Scott, Foresman & Co, 1975.

Demeurisse G, Demol O, Robaye E. Motor evaluation in vascular hemiplegia. EurNeurol. 19: 382-389, 1980.

DeSouza L, Langton-Hewer R, Miller S. Assessment of recovery of arm control in hemiplegic stroke patients. I. Arm function tests. Int J Rehab Med. 2: 3-9, 1980.

Dettmann M. Relationships among walking performance, postural stability, and functional assessments of the hemiplegic patient. Am J Phys Med. 65: 77-90, 1987.

DeVellis R. Scale Development: theory and applications. London: Sage Publications, 1991.

Deyo R, Centaur R. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis. 39: 897-906, 1986.

Deyo R, Inui T. Toward clinical applications of health status measures: sensitivity of scales to clinically important change. Health Serv Res. 19: 275-289, 1984.

Dietz V. Human neuronal control of automatic functional movements: interaction between central programs and afferent input. Physiol Rev. 72: 33-69, 1992.

Dove H, Schneider K, Wallace J. Evaluating and predicting outcome of acute cerebral vascular accident. Stroke. 15: 858-864, 1984.

Duncan P, Goldstein L, Horner R, Landsman P, Samsa G, Matchar D. Similar motor recovery of upper and lower extremities after stroke. Stroke. 25(6): 1181-1188, 1994.

Duncan P. Stroke: physical therapy assessment and treatment. In: Contemporary Management of Motor Control Problems: proceedings of the II Step Conference. Lister M. ed. Fredericksburg, VA: Bookcrafters, 1991.pp. 209-217.

Duncan P, Badke M. Stroke Rehabilitation: the recovery of motor control. Chicago, Il: Yearbook Medical Publishers, 1987.

Duncan P, Propst M, Nelson S. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Phys Ther. 63: 1606-1610, 1983.

Feigenson J, Fries J, Kunitz S. Time-oriented functional profile: practical applications in a stroke data-base model. Arch Phys Med Rehabil. 60: 512-516, 1979.

Feinstein A. Clinametrics. New Haven: Yale University Press, 1987.

Feinstein A, Cicchetti D. High agreement but low kappa: I.the problems of two paradoxes. J Clin Epidemiol. 43: 543-549, 1990.

Fleiss J, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as a measure of reliability. Ed Psych Msmt. 33: 613-619, 1973.

Fugl-Meyer A, Jaaski L, Legman I, Olsson S, Steglind S. The post-stroke Hemiplegic patient I. A method for evaluation of physical performance. Scand J Rehab Med. 7: 13-31, 1975.

Gans B, Haley S, Hallenborg S, Mann N. Description and interobserver reliability of the Tufts Assessment of Motor Performance. Am J Phys Med Rehabil. 67: 202 - 210. 1988.

Gowland C. Recovery of motor function following stroke: profile and predictors. Physiother Can. 34: 77-84, 1982.

Gowland C, Stratford P, Ward M, Moreland J, Torresin W. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. Stroke. 24: 58-63, 1993.

Gowland C, Torresin W, Ward M, Stratford P. Stroke Impairment and Disability Measure. Final Report to Program For Grant #6686-3771-45, 1992

Gowland C, Torresin W, Stratford P. Chedoke-McMaster Stroke Assessment: a comprehensive clinical and research measure. In: Proceedings of the 11th International Congress of the World Confederation for Physical Therapy. Barbican Centre, London. 2: 851-853, 1991.

Granger C. A conceptual model for functional assessment. In: Functional Assessment in Rehabilitation Medicine. Granger C, Gresham G.eds. Baltimore: Williams & Wilkins, 1984. pp. 14-25.

Grandger C, Gresham G. International Classification of Impairments, Disabilities and Handicaps (ICIDH) as a conceptual basis for stroke outcome research. Stroke. 21 (suppl II): II66-II67, 1990.

Gray C, French J, Bates D,Cartlidge N. Motor recovery following acute stroke. Age and Ageing. 19: 179-184, 1990.

Green L, Lewis F. Measurement and Evaluation in Health Education and Health Promotion. Los Angeles: Mayfield Publishing, 1986.

Gresham G, Fitzpatrick T, Wolf P. Residual disability in stroke survivors: the Framingham study. New Eng J Med. 293: 954-956, 1975.

Gresham G, Phillips T, Wolf P. McNamara P. Epidemiologic profile of a long term stroke disability: the Framingham study. Arch Phys Med Rehabil. 60: 487-491, 1979.

Grimby G, Finnstam J, Jette A. On the application of the WHO handicap classification in rehabilitation. Scand J Rehab Med. 20: 93-98, 1988.

Gronley J, Perry J. Gait analysis techniques: Ranch Los Amigos gait laboratory. Phys Ther. 64: 1831-1838, 1984.

Guarna F, Corriveau H, Chamberland J, Arsenault B. An evaluation of the hemiplegic subject based on the Bobath approach. Scand J Rehab Med. 20: 1-16, 1988.

Guccione A. Physical therapy diagnosis and the relationship between impairments and function. Phys Ther. 41: 10-15, 1991.
Guyatt G, Walter S, Normen G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Disabil. 40: 171-178, 1987.

Helmstadter G. Principles of Psychological Measurement. New York: Appleton-Century-Crofts, 1964.

Henley S, Pettits, Todd-Pokropec A, Tupper A. Who goes home? Predictive factors in stroke recovery. J Neurol, Neurosurg, and Psych. 48: 1-6, 1985.

Holden M, Gill K, Magliozzi M. Gait assessment for neurologically impaired patients: standards for outcome assessment. Phys Ther. 66: 1530-1539, 1986.

Horak F. Clinical measurement of postural control in adults. Phys Ther. 67: 1881-1885, 1987.

Huijbregts M. Content and construct validity of the disability inventory of the Chedoke-McMaster Stroke Assessment. Thesis for Master of Health Science, McMaster University, December, 1992.

Jebsen R, Taylor N, Trieschmann R, Trotter M, Howard L. An objective and standardized test of hand function. Arch Phys Med Rehabil. 50: 311-319, 1969.

Jebsen R, Trieschmann R, Mikulic M, Hartley R. Measurement of time in a standardized test of patient mobility. Arch Phys Med Rehabil. 53: 170-175, 1970.

Jerntorp P, Berglund G. Stroke registry in Malmo, Sweden. Stroke. 23: 357-361, 1992.

Jette A. Physical disablement concepts for physical therapy research and practice. Phys Ther. 74: 380-386, 1994.

Jette A. Concepts of health and methodological issues in functional assessment. In: Functional Assessment in Rehabilitation Medicine. Granger C, Gresham G.eds. Baltimore: Williams & Wilkins, 1984. pp. 46-64.

Jette A. State of the art in functional status assessment. In: Measurement in Physical Therapy. Rothstein J.ed. New York: Churchill Livingstone, 1985. pp. 137-168.

Joliffe I. Principal component analysis and exploratory factor analysis. Stat Meth Med Res. 1: 69-95, 1992.

Jongbloed L, Jones W. Prediction of recovery after stroke: an examination of outcome criteria. Can J Rehabil. 2: 87-92, 1988.

Kane R, Kane R. Assessing the Elderly. Toronto: Lexington Books, 1981.

Kazis L, Anderson J, Meenan R. Effect sizes for interpreting changes in health status. Med Care. 27: (supplement 3): S178-S190, 1989.

Keith R. Functional assessment measures in medical rehabilitation: current status. Arch Phys Med Rehabil. 65: 74-78, 1984.

Kelly-Hayes M, Wolf P, Kase C. Time course of functional recovery after stroke: the Framingham study. J Neurol Rehab. 3: 65-70, 1989.

Kirshner B, Guyatt G. A methodological famework for assessing health indices. J Chron Dis. 38: 27-35, 1985.

Kline P. A Handbook of Test Construction. London: Methuen, 1986.

Knutsson E, Martensson A. Dynamic motor capacity in spastic paresis and its relation to prime mover dysfunction, spastic reflexes and antagonist co-activation. Scand J Rehabil Med. 12:93-106, 1980.

Knutsson E, Richards C. Different types of disturbed motor control in gait of hemiparetic patients. Brain. 102: 405-430, 1979.

Kotila M. Declining incidence and mortality of stroke? Stroke. 15: 255-259, 1984.

Kramer M, Feinstein A. Clinical biostatistics: the biostatistics of concordance. Clin Pharmacol Ther. 29: 111-123, 1981.

Krebs D. Measurement theory. Phys Ther. 67: 1834-1839, 1987.

Kusoffsky A, Wadell I, Nilsson B. The relationship between sensory impairment and motor recovery in patients with hemiplegia. Scand J Rehab Med. 14: 27-32, 1982.

Landis J, Koch G. A review of statistical methods in the analysis of data arising from observer reliability studies (Parts I and II). Stat Neerlandica. 29: 101-123, 151-161, 1975.

LaRocca N. Statistical and methodological considerations in scale development. In: Quantification of Neurological Rehabilitation. Munsat T.ed. Boston: Butterworths Publishing, 1989. pp.49-68.

LaVigne J. Hemiplegia Sensorimotor Assessment Form. Phys Ther. 54: 128-134, 1974.

Law M. Measurement in Occupational Therapy: scientific criteria for evaluation. CJOT. 54: 133-138, 1987.

Leahy P. Motor control assessment. In: Motor Control and Physical Therapy: Theoretical Framework and Practical Applications. Montgomery P, Connolly B.eds. Hixson, TN: Chattanooga Group Inc., 1991. pp. 69-82.

Liang M, Larson M, Cullen K, Schwartz J. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis and Rheumatism. 28: 542-547, 1985.

Lindmark B, Hamrin E. Evaluation of functional capacity after stroke as a basis for active intervention. Scand J Rehabil Med. 20: 103-115, 1988.

Lincoln A, Leadbitter I. Assessment of motor function in stroke patients. Physiotherapy. 65: 48-51, 1979.

Loewen S, Anderson B. Reliability of the modified motor assessment scale and Barthel Index. Phys Ther. 68: 1077-108, 1988.

Loewen S, Anderson B. Predictors of stroke outcome using objective measurement sacales. Stroke. 21: 78-81, 1990.

Lyle R. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehab Research. 4: 483-492, 1981.

MacKenzie R, Charlson M, DiGioa D. Can the sickness impact profile measure change?: an example of scale assessment. J Chron Dis. 39: 429-438, 1986.

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

Mayo N. Epidemiology and recovery of stroke. In: Long-term Consequences of Stroke. Teasell R.ed. State of the Art Reviews in Physical Medicine. PA: Hanley & Belfus, 1993.

Mayo N, Hendlisz J, Goldberg M. Destinations of stroke patients discharged from Montreal area hospitals (1984-85). Stroke 20: 351-356, 1989.

Mayo N, Goldberg M, Levy A. Changing rates of stroke in the province of Quebec, Canada: 1981-1988. Stroke. 22: 590-595, 1991a.

Mayo N, Korner-Bitensky N, Becker R. Recovery time of independent function post stroke. Am J Phys Med Rehabil. 70: 5-12, 1991b.

McDowell I, Newell C. Measuring Health: a guide to rating scales and questionnaires. Oxford University Press, 1987.

Michels E. Measurement in physical therapy. Phys Ther.63: 209-215, 1983.

Moreland J, Gowland C, VanHullenaars S, Huijbregts M. Theoretical basis of the Chedoke-McMaster Stroke Assessment. Physiother Can. 45(4): 231-238, 1993.

Moskowitz E, Lightbody F, Freitag N. Long term follow-up of the post stroke patient. Arch Phys Med Rehabil. 53: 167-172, 1972.

Norman G. Issues in the use of change scores in randomized trials. J Clin Epidemiol. 42: 1097-1105, 1989.

Nunnally J. Psychometric Theory. New York: McGraw-Hill, 1978.

Olson T. Arm and leg paresis as outcome predictors in stroke rehabilitation. Stroke. 21: 247-251, 1990.

Osberg J, DeJong G, Haley S. Predicting long-term outcome among post-rehabilitation stroke patients. Am J Phys Med. 67: 94-103, 1988.

Ottenbacker K, Barrett K. Measures of effect size in reporting rehabilitation research. Am J Phys Med Rehabil. 68: 52-58, 1989.

Partridge C, Johnston M, Edwards S. Recovery from physical disability after stroke: normal patterns as a basis for evaluation. The Lancet. 373-375, 1987.

Podsiadlo D, Richardson S. The timed "up & go": a test of basic functional mobility for frail elderly persons. JAGS. 39: 142-148, 1991.

Poole J, Whitney S. Motor assessment scale for stroke patients: concurrent validity & interrater reliability. Arch Phys Med Rehabil. 69: 195-197, 1988.

Riddle D, Finucane S, Rothstein J. Intrasession and intersession reliability of hand-held dynamometer measurements taken on brain-damaged patients. Phys Ther. 69: 182-194, 1989.

Rosecrance J, Giuliani C. Kinematic analysis of lower-limb movement during ergometer pedaling in hemiplegic and nonhemiplegic subjects. Phys Ther. 71: 334-343, 1990.

Rosenbaum P, Russell D, Cadman D, Gowland C, Jarvis S. Issues in measuring change in motor function in children with cerebral palsy: a special communication. Phys Ther. 70: 125-131, 1990.

Rothstein J. Measurement and clinical practice: theory and application. In: Measurement in Physical Therapy. Rothstein J.ed. New York: Churchill Livingstone, 1985. pp.1-46.

Sackett D, Chambers L, MacPherson A, Goldsmith C, Mcauley R. The development and application of indices of health: general methods and a summary of results. AJPH 67: 423-428, 1977.

Sandin K, Smith B. The measure of balance in sitting in stroke rehabilitation prognosis. Stroke. 21: 82-86, 1990.

Sanford J, Moreland J, Swanson L, Stratford P, Gowland C. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. Phys Ther. 73(7): 447-454, 1993.

SAS Institute Inc. SAS Procedures Guide for Personal Computers. Version 6 edition, 1985.

Schmidt E, Smirnov V, Ryabova V. Results of the seven year prospective study of stroke patients. Stroke. 19: 942-949, 1988.

Seaby L, Torrance G. Reliability of a physiotherapy functional assessment used in a rehabilitation setting. PhysTher Can. 41: 264-269, 1989.

Seaby L. Mobility: A Functional Domain for Clinical Assessment and Program Evaluation in Physiotherapy. Master's thesis, McGill University, Montreal, 1987.

Shah S, VanclayF, Cooper B. Predicting discharge status at commencement of stroke rehabilitation. Stroke. 20: 766-769, 1989.

Shumway-Cook A, Horak F. Assessing the influence of sensory interaction on balance. Suggestions from the field. Phys Ther. 66: 1548-1550, 1986.

Sjostrom M, Fugl-Meyer A. Post-stroke hemiplegia: crural muscle strength and structure. Scand J Rehabil Med. Supplement 7:53-61, 1981.

Soeken K, Prescot P. Issues in the use of kappa to estimate reliability. Med Care. 24: 733-741, 1986.

Statistics Canada. Hospital morbidity, Catalogue 82-206, 1989.

Streiner D, Norman G. Health Measurement Scales: a practical guide to their development and use. Oxford: Oxford University Press, 1991.

Task Force on Standards for Tests and Measurements in Physical Therapy Practice. Phys Ther. 71: 589-622, 1991.

Task Force on Stroke. Symposium recommendationsfor methodology in stroke outcome research. Stroke. 21(suppl II): II68-II73.

Tinetti M. Performance oriented assessment of mobility problems in elderly patients. JAGS. 34: 119-126, 1986.

Tripp E, Harris S. Test-retest reliability of isokinetic knee extension and flexion torque measurements in persons with spastic hemiparesis. Phys Ther. 71: 390-396, 1991.

Twitchell T. The restoration of motor function following hemiplegia in man. Brain. 74: 443-480, 1951.

Wade D. Stroke: rehabilitation and long-term care. Lancet. 339(8796): 793-795, 1992.

Wade D, Hewer R. Functional abilities after stroke: measurement, natural history and prognosis. J Neurol Neurosurg Psychiatry. 50: 177-182, 1987.

Wade D, Wood V, Heller A. Walking after stroke: measurement and recovery over the first 3 months. Scand J Rehab Med. 19: 25-30, 1987.

Wade D, Wood V, Hewer R. Recovery after stroke: the first 3 months. J Neurol Neurosurg Psychiatry. 48: 7-13, 1985.

Walker A, Robins M, Weinfield F. Chapter 3: Clinical findings in Weinfeld FD: the national survey of stroke. Stroke. 12: 13-44, 1981.

Ware J. Standards for validating health measures: definition and content. J Chron Dis. 40: 473-480, 1987.

Watkins M, Harris B, Kozlowski B. Isokinetic testing in patients with hemiparesis: a pilot study. Phys Ther. 64: 184-189, 1984.

Weiner E, Stewart B. Assessing Individuals. Boston: Little-Brown, 1984.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1: 80-89, 1945.

Winstein C. Motor learning considerations in stroke rehabilitation. In: Stroke Rehabilitation: recovery of motor control. Duncan P, Badke M.eds. Chicago: Yearbook Medical Publishers, 1987.

Wood-Dauphinee S, Williams I, Shapiro S. Examining outcome measures in a clinical study of stroke. Stroke. 21: 731-739. 1990.

Wood-Dauphinee S, Williams I. Reintegration to normal living as a proxy to quality of life. J Chron Dis. 40(6): 491-499, 1987.

Working Group on Outcome Measures in Physiotherapy. Health and Welfare Canada Contract. 1992. Information requests: Nancy Mayo, School of Physial and Occupatipnal Therapy, McGill University, Montreal, Quebec.

World Health Organization (WHO) International Classification of Impairments, Disabilities, and Handicaps. Geneva: World Health Organization, 1980.

World Health Organization: Recommendations on stroke prevention, diagnosis, and therapy. Stroke. 20: 1407-1431, 1989.

Zeller R, Carmines E. Measurement in the Social Sciences: the link between theory and data. London: Cambridge University Press. 1980.

APPENDICES

APPENDIX 1
The Original STREAM

STroke REhabilitation Assessment of Movement (STREAM)

The Steps to Recovery

General Instructions for Using the STREAM Assessment

1. The patient should be in his/her usual state of attention and health.
2. Instructions should be explained, demonstrated and repeated to the patient as necessary.
3. Therapists may use their judgement as to the order in which the items are tested.
4. If necessary, the patient is permitted two attempts on each item.
5. Patients may be given verbal encouragement and reassurance.
6. If the assessment must be interrupted for any reason then it may be restarted from where it was left off (within a 24 hour period)
7. The following equipment should be on hand:

   - armless chair of standard height or plinth
   - stairs
   - small towel.

8. Unless otherwise stated the starting position is sitting on a straight armless chair (or plinth) with feet flat on the floor and the affected hand is placed in the lap.
9. The items evaluating the upper extremity, lower extremity and sitting balance are scored:

   0 if the patient is unable to carry out the activity
   1 if the patient is able to carry out the activity

10. The items evaluating standing balance and walking are scored:

    0 if the patient is unable to carry out the activity
    1 if the patient is able to carry out the activity with assistance of another person (with or without an aid)
    2 is able to carry out the activity with the help of an aid or if holding on to a rigid support
    3 is able to carry out the activity independently and safely without the help of an aid.

Instructions for Use of STREAM

Upper Extremity

1. Affected hand touches ipsilateral ear

   1 elbow must be abducted from the side of the body by at least 45° and the patient must keep the head and trunk in the midline

2. Affected hand touches sacrum

   1 patient must be able to carry out the activity without bending the trunk

3. Raises affected arm to **forward**-horizontal flexion

   1 range of permissible error is ± 20° ie shoulder flexion can range from 70°-110°, deviation in the horizontal plane can range from 20° adduction to 20° abduction, and the elbow must not be flexed more than 20°

4. Raises affected arm to **lateral**-horizontal flexion

   1 range of permissible error is ± 20° ie abduction can range from 70°-110°, and the elbow must not be flexed more than 20°.

5. Raises affected arm to **full** forward flexion

   1 the arm must be raised above the horizontal position and the elbow must not be flexed more than 20°

6 Supination and pronation with elbow **flexed** to 90°

   1 movement must occur through more than 1/3 of available range and the elbow must remain tucked in at the side of the body

7. Actively closes the affected hand

   1 active movement of the fingers must be observed

8. Actively opens the affected hand

   1 active movement of the fingers must be observed

9. Touches each fingertip to thumb of affected hand

Lower Extremity

10. Flexes hip and knee while lying supine

    1 active movement through at least 1/3 of available range (heel does not have to leave the surface)

11. Actively extends affected knee while sitting

    1 patient must be able to carry out the activity without leaning back or extending the hip, active movement must occur through 2/3 of available range

12. Active dorsiflexion of affected ankle while sitting

    1 the ball of the foot must be lifted and the heel must remain on the floor

13. Active dorsiflexion of affected ankle with knee extension while standing

    1 the ball of the foot must be lifted, the knee must not flex more than 10° and the heel must remain on the floor

14. Flexion of affected knee with hip extension while standing

    1 hip must remain in the neutral position (less than 10° flexion) and knee must be flexed beyond 45°

15. Abduction of affected hip with knee extension while standing

    1 active movement must occur through at least 1/3 of available range, hip must remain in the neutral position (less than 10° flexion) and the knee must retain near full extension (less than 10% flexion)

Sitting Balance

16. Maintains erect sitting position for one minute

17. Maintains sitting balance while raising folded arms to forehead

    1 arms are folded across chest, the **unaffected** arm supporting the **affected** arm, patient raises forearms to touch forehead without bending the head forward

18. Maintains sitting balance while moving folded arms side to side

    1 arms are folded across chest, the **unaffected** arm supporting the **affected** arm, with the arms raised to clear chest, the patient moves arms from side to side while maintaining the head facing forward, at each movement the elbow approaches the mid-line

19. Bends forward to touch floor with **unaffected** hand and returns to the sitting position.

Standing and Walking

Note - for items 20 to 34 assign:

0 if the patient is unable to carry out the activity
1 if the patient is able to carry out the activity with assistance of another person (with or without an aid)
2 is able to carry out the activity with the help of an aid or if holding on to a rigid support
3 is able to carry out the activity independently and safely without the help of an aid.

20. Rises to standing from sitting

21. Maintains standing for one minute

22 Slides **unaffected** leg 12" forwards then backwards

1,2,3activity must be accomplished in one smooth movement, to reduce friction a small towel may be placed under the foot

23. Lifts **unaffected** foot off the ground

1,2,3the whole foot must leave the ground at least momentarily

24. Turns 90^0 to **unaffected** side

25. Turns 90^0 to **affected** side

26. Takes 3 steps forwards

1,2,3one step is accomplished by advancing one foot and bringing the other foot to meet the first

27. Takes 3 steps backwards

1,2,3one step is accomplished by moving one foot back and bringing the other foot to meet the first

28. Takes 3 steps sideways to **unaffected** side

1,2,3one step is accomplished by moving one foot and bringing the other foot to meet the first

29. Takes 3 steps sideways to **affected** side

1,2,3one step is accomplished by moving one foot and bringing the other foot to meet the first

30. Walks **25'**

1,2,3walks on a smooth obstacle free surface

31. Walks **up** 3 stairs **without alternating** feet

1,2,3the two feet <u>are</u> on the same step together
Note use of the handrail counts as using an aid (score 2)

32. Walks **down** 3 stairs **without alternating** feet

1,2,3the two feet <u>are</u> on the same step together
Note use of the handrail counts as using an aid (score 2)

33. Walks **up** 3 stairs **alternating** feet

1,2,3the two feet <u>are never</u> on the same step together
Note use of the handrail counts as using an aid (score 2)

34. Walks **down** 3 stairs **alternating** feet

1,2,3the two feet <u>are never</u> on the same step together
Note use of the handrail counts as using an aid (score 2)

APPENDIX 2

Documents from The Content Verification Survey

CONTENT VERIFICATION SURVEY

Dear colleague:

We are in the process of developing an instrument to assess motor functioning of stroke patients, the **STroke REhabilitation Assessment of Movement (STREAM)**. This is a collaborative effort between Dr. Nancy Mayo, Director of Research at the Jewish Rehabilitation Hospital in Montreal, Dr. Sharon Wood-Dauphinee, Director School of Physical and Occupational Therapy McGill University, and myself, Kathy Daley BSc PT, masters student in Rehabilitation Science at McGill University. As a thesis project I will be evaluating the measurement properties of STREAM (ie. the reliability, validity, and responsiveness) in order to assess STREAM's potential for use as an outcome measure for stroke rehabilitation. The first step in this project is to verify the content of the scale, and for this task your expertise is needed.

Background of STREAM:

STREAM has been designed to be used by physiotherapists to monitor patients' motor function over time, and for evaluating the effectiveness of physiotherapy interventions in stroke rehabilitation. The items which make up the STREAM cover a wide range of movement patterns that are typically assessed by physiotherapists during rehabilitation following stroke. The items have been adapted from several existing assessment scales for motor function, and from the clinical experience of the scale developers. STREAM differs from existing motor assessments in that it covers a wide scope of motor components of interest to physiotherapists, yet is brief and easy to administer and score. The scoring of STREAM is intended to be objective so that a high reliability or reproducibility of scores can be achieved. The scale is designed for the evaluation of stroke patients with motor deficits ranging from mild to severe.

This survey is intended to elicit feedback from physiotherapists who have experience working with stroke patients. Your opinions regarding the potential usefulness of this measurement instrument will be greatly appreciated and will facilitate the process of scale development. Please feel free to include any suggestions as to modifications or refinements which may improve the STREAM.

A copy of STREAM and instructions for use are enclosed and should be read prior to completing the survey regarding the contents of STREAM

Thank you in advance for taking the time to complete this brief survey and for your input!

STroke REhabilitation Assessment of Movement
STREAM: The Steps to Recovery

Please rate each items as to its importance using the following key:

1 crucial
2 very important
3 moderately important
4 neither important nor unimportant
5 unimportant

	Rating	Comments
Upper Extremity		
1. Affected hand touches ipsilateral ear	1 2 3 4 5	_____
2. Affected hand touches sacrum	1 2 3 4 5	_____
3. Raises affected arm to forward-horizontal flexion	1 2 3 4 5	_____
4. Raises affected arm to lateral-horizontal flexion	1 2 3 4 5	_____
5. Raises affected arm to full forward flexion	1 2 3 4 5	_____
6. Supination and pronation	1 2 3 4 5	_____
7. Actively closes the affected hand	1 2 3 4 5	_____
8. Actively opens the affected hand	1 2 3 4 5	_____
9. Touches each fingertip to thumb of affected hand	1 2 3 4 5	_____
Lower Extremity		
10. Flexes hip and knee while lying supine	1 2 3 4 5	_____
11. Extends affected knee while sitting	1 2 3 4 5	_____
12. Dorsiflexion of affected ankle in sitting	1 2 3 4 5	_____
13. Dorsiflexion of affected ankle with knee extension while standing	1 2 3 4 5	_____
14. Flexion of affected knee with hip extension while standing	1 2 3 4 5	_____
15. Abduction of affected hip with knee extension while standing	1 2 3 4 5	_____

Sitting Balance

16. Maintains sitting position 1 minute.	1 2 3 4 5	_____
17. Maintains sitting balance raising folded arms to forehead	1 2 3 4 5	_____
18. Maintains sitting balance moving folded arms side to side	1 2 3 4 5	_____
19. Bends forward to touch floor with unaffected hand and returns to sitting	1 2 3 4 5	_____

Standing Balance

20. Rises to standing from sitting	1 2 3 4 5	_____
21. Maintains standing for one minute	1 2 3 4 5	_____
22. Slides unaffected leg 12" forwards then backwards	1 2 3 4 5	_____
23. Lifts unaffected foot	1 2 3 4 5	_____

Walking

24. Turns 90^0 to unaffected side	1 2 3 4 5	_____
25. Turns 90^0 to affected side	1 2 3 4 5	_____
26. Takes 3 steps forwards	1 2 3 4 5	_____
27. Takes 3 steps backwards	1 2 3 4 5	_____
28. Takes 3 steps sideways to unaffected side	1 2 3 4 5	_____
29. Takes 3 steps sideways to affected side	1 2 3 4 5	_____
30. Walks 25'	1 2 3 4 5	_____
31. Walks up 3 stairs without alternating feet	1 2 3 4 5	_____
32. Walks down 3 stairs without alternating feet	1 2 3 4 5	_____
33. Walks up 3 stairs alternating feet	1 2 3 4 5	_____
34. Walks down 3 stairs alternating feet	1 2 3 4 5	_____

Please answer the following by placing a checkmark in the appropriate space provided and add your comments where requested (THANK YOU!):

Therapist profile

A) Number of stroke patients assessed in the past year: <10 ___ 10-30___ >30___

B) Years of experience working with stroke patients: <1yr___ 1-5___ >5yr___

C) Current service area:

i) staff therapist___ senior___ clinical specialist___ other_____

ii) acute___ rehabilitation___ long-term care___ private practice___ out-patient___ day-hospital___ home-care___ other_____

iii) general (mixed caseload)___ mixed neuro ___ stroke service___ other_____

D) Name of facility employed at: _____

COMMENTS on STREAM

1) Considered collectively, is the total group of items adequate to evaluate the motor function of stroke patients? Yes___ No___

If not, what would you add?_____

_____

2) Is each item clear and easy to understand? Yes___ No___

List items with problems (by number--please comment on problem)

_____

_____

_____

_____

_____

_____

_____

_____

_____

3) Is the scoring (able / unable for limbs and sitting balance; 4 point scale for standing items depending on degree of assistance required) appropriate? Yes___ No___

If no, how might the scoring be improved?

_____

_____

_____

4) Do you agree that this scale will show differences in motor function among persons undergoing physiotherapy interventions following stroke?

Agree strongly___ Agree___ Disagree___ Disagree strongly___

5) Would you use this scale in your clinical practice to evaluate patient outcome (following evaluation of reliability and validity of the scale) ? Yes___ No___

If not, why not?

_____

_____

_____

_____

_____

_____

6) Are there any other comments you would like to make regarding STREAM?

THANK YOU ONCE AGAIN FOR YOUR TIME AND INPUT!

Please return to: Attn: Kathy Daley (Masters student)
School of P&OT, McGill University
3654 Drummond Street
Montreal, Que. H3G 1Y5

APPENDIX 3

Documents from The Consensus Panel Meetings

APPENDIX 3.1.1
Cover Letter to Panel One Participants

Research Department, Jewish Rehabilitation Hospital
3205 Place Alton Goldbloom
Chomedy, Laval, Que. H7V 1R2

February 13, 1993

Dear colleague:

We are in the process of developing and evaluating an instrument to assess the motor function of stroke patients, the **STroke REhabilitation Assessment of Movement (STREAM)**. This is a collaborative effort between Dr. Nancy Mayo, Epidemiologist at the Jewish Rehabilitation Hospital in Montreal, Dr. Sharon Wood-Dauphinee, Associate Dean School of Physical and Occupational Therapy McGill University, and Kathy Daley BSc PT, masters student in Rehabilitation Science at McGill University.

We are recruiting experienced therapists from the Montreal area to participate in a concensus group to refine the STREAM. We are asking for your assistance in identifying potential participants. Each participant will be asked to review the STREAM, and to attend one meeting (to be held on an evening in early February). For this effort each participant will receive a small honorarium.

We have enclosed some information on the background of STREAM, and information for potential concensus group participants. Could you please bring this project to the attention of your staff, and ask any interested people to contact Susie Rosenmeier at the Research Department of the Jewish Rehabilitation Hospital. If you need any further information, please don't hesitate to call.

Thank you for your assistance on this project !

Attn: Susie Rosenmeier,
Research Department,
Jewish Rehabilitation Hospital
688-9550, ext (437)

BACKGROUND OF STREAM

The STroke REhabilitation Assessment of Movement (STREAM), developed in 1986 at the Jewish Rehabilitation Hospital (JRH), was designed to be used by physiotherapists to monitor patients' motor function over time, and for evaluating the effectiveness of physiotherapy interventions in stroke rehabilitation. The items which make up the STREAM were contributed by physical and occupational therapists and members of the research department, and cover a wide range of movement patterns that are typically assessed by physiotherapists during rehabilitation following stroke. The items were adapted either from the clinical experience of the scale developers or from existing published assessments (Bobath, 1978; Brunnstrom, 1970). STREAM was designed to be a relatively comprehensive assessment of motor function suitable for evaluating stroke patients with motor deficits ranging from mild to severe. It includes a total of 34 items (of graded difficulty) assessing upper and lower limb movement patterns, basic mobility, and balance. The 15 items assessing limb movement patterns (6 for the lower extremity and 9 for the upper extremity) and 4 items performed in sitting are scored on a two point ordinal scale (able/unable); 6 items performed in standing and 9 mobility items are scored on a four point ordinal scale ranging from 0:unable to 3:independent without aids or assistance. The scoring of STREAM is intended to be objective so that a high reliability or reproducibility of scores can be achieved; it is also brief and easy to administer. The scale has been piloted for clinical utility at the JRH for a number of years and has proven clinically acceptable.

INFORMATION FOR POTENTIAL CONCENSUS GROUP PARTICIPANTS

We are evaluating the content validity of STREAM, and are recruiting experienced therapists (from all phases of stroke rehabilitation, with a minimum of one year of experience working with stroke patients) to participate in a concensus group to refine the STREAM. Through the suggestions made by the group, we hope to achieve concensus on the changes that should be made, and to develop an instrument that is useful to physical therapists for measuring motor function following stroke. Your participation would involve the following

1) we would ask you to review the STREAM (prior to the meeting) and think about changes you would make to improve the instrument (additions, deletions, clarifications, and scoring of items), and

2) we would ask you to attend a meeting (of approximate 3 hours duration, to be held on an evening in early February) to discuss changes that should be made to refine the STREAM. At this meeting you will be shown a videotaped patient assessment on STREAM, to further familiarize you with the instrument, prior to the discussion.

For your participation, you would receive a small honorarium. Refreshments will be provided on the evening of the "think tank".

If you would be interested in participating, please contact:

Susie Rosenmeier,
Research Department, JRH
688-9550, ext 437

APPENDIX 3.1.2

Please use this form to list the changes that you would recommend to improve the STREAM, in terms of item additions, deletions, modifications, clarifications, and scoring. Your list should be compiled prior to the meeting, so that we can discuss the ideas as a group. You may include as many or as few ideas or details as you feel are important. We welcome any suggestions that will further improve the STREAM, and make it an instrument that will be widely accepted for use in clinical practice and in research.

SUGGESTED CHANGES to STREAM

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____
25. _____
26. _____
27. _____
28. _____
29. _____
30. _____

(use other side if additional space required)

APPENDIX 3.2.1
Cover Letter to Panel Two Participants

Dear concensus panel participant:

Thank you for agreeing to participate in this project! We are in the process of developing and evaluating an instrument to assess the motor function of stroke patients, the **STroke REhabilitation Assessment of Movement (STREAM)**. This is a collaborative effort between Dr. Nancy Mayo, Epidemiologist at the Jewish Rehabilitation Hospital in Laval, Dr. Sharon Wood-Dauphinee, Associate Dean School of Physical and Occupational Therapy McGill University, and Kathy Daley BSc PT, masters student in Rehabilitation Science at McGill University. We are evaluating the content validity of STREAM, and are recruiting experienced therapists (from all phases of stroke rehabilitation, with a minimum of one year of experience working with stroke patients) to participate in a concensus group (the second in a series of panels) to refine the STREAM. Through the suggestions made by the group, we hope to achieve concensus on the changes that should be made and to develop an instrument that is useful to physical therapists for measuring motor function following stroke. You were recommended to us by members of the first panel due to your experience and expertise working with stroke patients. Your participation would involve the following:

1) we would ask you to review the STREAM and think about changes you would make to improve the instrument (additions, deletions, clarifications, and scoring of items), and

2) we would ask you to attend a meeting (of approximate 3 hours duration) to discuss changes that should be made to refine the STREAM. At this meeting you will be shown a videotaped patient assessment on STREAM, to further familiarize you with the instrument, prior to the discussion.

For your participation, you would receive a small honorarium. Refreshments will be provided on the evening of the "think tank".

To familiarize yourself with the STREAM you might consider assessing a few patients using STREAM As you review the STREAM, make note of any changes that you would recommend to further improve the STREAM (on the form provided). We have enclosed:

a) a brief description of the purpose and scope of STREAM
b) a copy of STREAM
c) a form for you to write your suggestions on (please bring this to the meeting)
d) a meeting agenda

If you have any questions, please don't hesitate to call (questions should be directed to Susie Rosenmeier, Research Department, Jewish Rehabilitation Hospital, **688-9550, ext 437**).

We look forward to seeing you on March 25th !

Note: Please confirm your intention to participate by contacting Susie. Thank you!

APPENDIX 3.2.2
First Revision of Stream

STroke REhabilitation Assessment of Movement (STREAM)

General Instructions for Using the STREAM

1. The patient should be in his/her usual state of attention and health

2. The patient should be dressed in clothing that does not restrict movement, and that allows the therapist to observe the movement clearly (eg. shorts and T-shirt). Comfortable walking shoes (or the patients usual footwear) should be worn when testing the activities performed in standing.

3. Instructions (*italicized*) should be given, demonstrated and repeated to the patient as necessary For the items testing voluntary movement of the limbs, ask the patient to perform the movement one time with unaffected side (to observe the available range and normal movement pattern, as well as the patient's comprehension of the test item). When necessary, therapists may assist the patient to maintain standing to allow the performance of movements of the unaffected lower extremity that require weightbearing through the affected limb (lower extremity items #5-7).

4. Patients may be given verbal encouragement and reassurance.

5. Therapists may assist the patient to achieve the starting positions specified, however, during the testing of each individual item no support (except as stated above) or physical assist should be provided

6. If necessary, the patient is permitted three attempts on each item and the best performance recorded.

7. Therapists may use their judgement as to the order in which the items are tested.

8. If the assessment must be interrupted for any reason, it may be restarted from where it was left off if done so within a 24 hour period.

9. The following equipment should be on hand:

-**sturdy stool** (or treatment plinth or armless chair) of a height such that patient can sit comfortably on a firm support with feet resting on the floor or on a small foot stool, with the hips and knees at 90^{o}

-**support surface** (firm, large enough to permit rolling safely, and raised approximately 1/2 meter off the ground); if using the patient's bed, it must be fully flat with encumbering bedding removed; alternatively, a large treatment plinth (raised mat) may be used

-**pillow**

-**stairs** with railings (departmental steps or full flight: standard height approximately 18 cm)

10. The item should be excluded (and test scored out of 100) if movement is limited due to marked restriction of passive range or due to pain.

I. VOLUNTARY MOVEMENT OF THE LIMBS

Scoring: **0** **unable** to perform the test movement through any visible range (includes flicker)

1 able to perform **partial** range of movement, but requires assistance, stabilization or support to complete or performs full movement with **marked deviation** from normal pattern

2 **completes** the movement in a manner that is **comparable to the unaffected side**

X activity not tested (specify why)

UPPER EXTREMITY

Starting position for items 1-12: sitting on support surface with feet flat on the floor (or on a small foot stool) with hips and knees at 90°, and with both hands resting on a pillow placed on the patient's lap with palms down.

1. SHRUGS SHOULDERS (SCAPULAR ELEVATION)
   "Shrug your shoulders as high as you can"

2. RAISES HAND TO TOUCH FOREHEAD
   "Raise your hand to touch your forehead"

3. PLACES HAND ON SACRUM
   "Reach behind your back and as far across toward the other side as you can"

4. RAISES ARM FORWARD TO 90° OF FORWARD FLEXION
   "Reach your arm straight out in front of you to horizontal"

5. RAISES ARM OVERHEAD TO FULLEST ELEVATION
   "Reach your hand as high as you can over your head, keeping your elbow straight"

6. RAISES ARM SIDEWAYS TO 90° OF ABDUCTION
   "Lift your arm sideways to horizontal, keeping your elbow straight"

7. SUPINATES AND PRONATES FOREARM WITH ELBOW FLEXED AT 90°
   "Keep your elbow bent and close to your side. Turn your forearm over so that your palm faces up, then turn your forearm over so that your palm faces down"

8. CLOSES HAND FROM FULLY OPENED POSITION
   "Make a fist, keeping your thumb on the outside"
   (Note: Must extend wrist to obtain full marks.)

9. OPENS HAND FROM FULLY CLOSED POSITION
   "Now open your hand all the way"

10. EXTENDS WRIST AND FINGERS
    "Lift your hand and straighten your fingers"

11. OPPOSES THUMB TO INDEX FINGER (tip to tip)
    "Make a circle with your thumb and index finger"

12. OPPOSES THUMB TO LITTLE FINGER (tip to tip)
    "Make a circle with your thumb and little finger"

LOWER EXTREMITY

SUPINE

1. FLEXES HIP AND KNEE IN SUPINE SO THAT FOOT IS FLAT ON BED (ATTAINS HALF-CROOKLYING POSITION)
    "Bend your hip and knee so that your foot rests flat on the bed"

SITTING: feet flat on the floor (or supported on a small foot stool); hips and knees at 90°

2. FLEXES HIP IN SITTING
    "Lift your knee as high as you can"

3. EXTENDS KNEE IN SITTING
    "Straighten your knee by lifting your foot up"

4. DORSIFLEXES ANKLE IN SITTING
    "Keep your heel on the ground and lift your toes off the floor as far as you can"

STANDING: holding onto a stable support (to assist balance) for items 5-7

5. ABDUCTS AFFECTED HIP WITH KNEE EXTENDED IN STANDING
    "Keep your knee straight and your hips level, and raise your leg to the side"

6. FLEXES AFFECTED KNEE WITH HIP EXTENDED IN STANDING
    "Keep your hip straight and bend your knee back, bringing your heel towards your bottom"

7. DORSIFLEXES AFFECTED ANKLE IN STANDING WITH KNEE EXTENDED
    "Keep your heel on the ground and lift your toes off the floor as far as you can"

II. BASIC MOBILITY

Scoring: **0** **unable** to perform the test movement

1 **able to perform part** of the movement actively, but requires **supervision** or **assistance** from an **aid** or an **individual** to complete or performs any or all of the movement with **marked deviation from normal**

2 **completes** the movement **independently and safely**

X *activity not tested (specify why)*

CROOKLYING

1. RAISES HIPS OFF BED IN LYING WITH KNEES BENT (BRIDGING)
   "Lift your hips as high as you can"
   Note: May be assisted to attain starting position (crooklying)

SUPINE

2. ROLLS TOWARDS AFFECTED SIDE
   "Roll onto your weak side"
   Note: Must achieve full side lying (pelvis perpendicular to the bed)

3. ROLLS TOWARDS UNAFFECTED SIDE
   "Roll onto your strong side"
   Note: Must achieve full side lying (pelvis perpendicular to the bed)

4. MOVES FROM LYING SUPINE TO SITTING UPRIGHT WITH FEET ON THE FLOOR
   (to either side; any functional and safe method)
   "Sit up and place your feet on the floor"

SITTING

5. REACHES FORWARD TO TOUCH UNAFFECTED FOOT WITH UNAFFECTED HAND AND RETURNS TO SITTING
   "With your strong hand, reach down and touch your foot on the same side"

6. RISES TO STANDING FROM SITTING
   "Stand up; try to take equal weight on both legs"
   Note: May push up with hand(s)

STANDING AND WALKING

7. TAKES 3 STEPS FORWARDS
   (Placing one foot in front of the other)
   "Take three average sized steps forwards, placing one foot in front of the other"

8. TAKES 3 STEPS BACKWARDS
   (Placing one foot behind the other)
   "Take three average sized steps backwards, placing one foot behind the other"

9. TAKES 3 STEPS SIDEWAYS TO UNAFFECTED SIDE
(Steps must be at least 25 cm wide)
"Take three average sized steps towards your strong side"

10. TAKES 3 STEPS SIDEWAYS TO AFFECTED SIDE
(Steps must be at least 25 cm wide)
"Take three average sized steps towards your weak side"

11. MOVES FEET TO TURN BODY 90° TOWARDS UNAFFECTED SIDE
"Take steps to turn on the spot towards your strong side"

12. MOVES FEET TO TURN BODY 90° TOWARDS AFFECTED SIDE
"Take steps to turn on the spot towards your weak side"

13. WALKS 10 METERS INDOORS, ON A SMOOTH OBSTACLE FREE SURFACE (WITHIN 20 SECONDS)
"Walk in a straight line over to ... (a specified point 10 meters away) "
Note: use of an orthotic counts as using an aid (score 1)

14. WALKS UP 3 STAIRS ALTERNATING FEET
"Walk up three stairs; place only one foot at a time on each step if you can"
Note: use of the handrail counts as using an aid (score 1)

15. WALKS DOWN 3 STAIRS ALTERNATING FEET
"Walk down three stairs; place only one foot at a time on each step if you can"
Note: use of the handrail counts as using an aid (score 1)

APPENDIX 3.2.3
Intent and Scope of Stream

STROKE REHABILITATION ASSESSMENT OF MOVEMENT (STREAM)

BACKGROUND OF STREAM

The Stroke Rehabilitation Assessment of Movement (STREAM) was developed in 1986 at the Jewish Rehabilitation Hospital (JRH), Chomedy-Laval, Quebec. The items included in the original STREAM were contributed by physical and occupational therapists and members of the research department; the items were adapted from clinical experience and from existing published assessments (Bobath, 1978; Brunnstrom, 1970). The scale has been piloted for clinical utility at the JRH for a number of years and has proven clinically acceptable.

INTENDED PURPOSE & SCOPE

The purpose of the STREAM is to provide a standardized, objective and quantitative assessment for measuring the motor recovery of stroke patients. It was designed to be used by physiotherapists for **monitoring motor recovery**, and for **evaluating the impact of therapeutic interventions** (such as physical therapy, medical and pharmacologic interventions, etc.) on the motor recovery of stroke patients.

The STREAM is intended for use as an **outcome measure** (for evaluating treatment outcomes), and for monitoring motor improvement-- as opposed to being an assessment tool for diagnosis and treatment planning. This is an important point, because the intended purpose of the instrument influences what items must be included and the amount of detail required. Although STREAM is not intended for planning treatment strategies, if used in isolation, it may be used in conjunction with measures of spasticity, ROM, sensation, balance, functional independence in ADL, etc. to obtain a detailed objective profile of a patient's physical function.

In the present form of the STREAM, the items included cover a wide range of motor activities that are typically assessed by physiotherapists for the purpose of evaluating the motor status of stroke patients. STREAM was designed to be a relatively comprehensive assessment, suitable for evaluating stroke patients with motor deficits ranging from mild to severe. The test items represent gradations of difficulty, to allow discrimination between patients with differing levels of motor impairment, and to monitor an individual's motor recovery over time. The items currently included in the instrument assess **upper and lower limb movement patterns** and **basic mobility**. The items assessing limb movement patterns are intended to provide a profile of basic motor status (motor recovery in the limbs); the remaining items should give a more functional picture of motor recovery, assessing simple motor tasks which require the integrated

movement of trunk and limbs. To date, the items included in the STREAM have been limited to simple movements in an attempt to minimize the potential confounding effects that variables other than motor recovery, such as cognition/perception and communication, have on the measurement of motor status.

The STREAM currently includes a total of 34 items, with 19 items assessing limb movement patterns (7 for the lower extremity and 12 for the upper extremity) and 15 items assessing basic mobility. A simple three point ordinal scoring system is employed; a patient is scored **0** if unable to perform the test movement, **1** if a movement can be partially completed but assistance or supervision is required or if the movement is performed with marked deviation from a normal pattern (as compared to unaffected limb where possible), and **2** if able to complete the movement in a manner that is qualitatively and quantitatively near normal (see scale for specific scoring directives). The scoring of STREAM is designed to be objective and quantitative so that a high reliability or reproducibility of scores should be achievable. STREAM is also brief (10-15 minutes) and easy to administer.

CONCEPTUAL FRAMEWORK

In devising a clinical instrument to measure motor recovery, a number of aspects must be considered, such as: the domains to be included, the methods by which these are assessed, and the uses and populations for which the measure is appropriate. Thus, motor recovery must be operationally and conceptually defined. However, given the complexity of motor performance in terms of the number of factors that potentially influence motor recovery and the functional performance of stroke patients, defining and measuring this phenomenon is a challenge. What the STREAM endeavors to measure has been referred to by a variety of terms, including: motor recovery, motor ability, motor function, functional motor recovery, motor performance, motor control, motricity, and motor output. Unfortunately, many of these terms are not well defined and may imply radically different concepts to different people. We decided to call the concept measured by the STREAM "voluntary movement and basic mobility", and the following conceptual framework was developed.

Motor recovery implicates a number of neurophysiologic processes (at the cellular or histochemical level), and a multitude of contributing factors such as: the size and side of the lesion, age, comorbid conditions, motivation, communication, cognition, and secondary impairments like contractures or oedema. It is very difficult to separate out the influences that these various processes and factors have on motor recovery, however, fundamentally, motor recovery is manifested by the re-emergence of voluntary movement and basic mobility-- that

is, what we observe as a result of motor recovery is improvement in a patient's ability to move. The STREAM is intended to measure these fundamental building blocks that reflect motor recovery, and that can be easily measured in the clinical setting.

In the context of physical rehabilitation following stroke (using the terminology of the World Health Organizations International Classification of Impairments, Disabilities, and Handicaps), impairments refer to the actual deficits observed (ie. the signs and symptoms, such as paresis, sensory deficits and abnormal tone which contribute to, or cause, motor dysfunction), whereas disability refers to the functional consequences of the impairment (such as reduced ability to perform purposeful movements, transfers, walking, and activities of daily living). The STREAM falls into the category of disability assessments: it is related hierarchically to other disability measures in that it measures basic motor disability-- one step beyond the level of the primary impairments, and one step before functional mobility and ADL measures.

EXPECTED CONTRIBUTIONS

The STREAM should: 1) streamline and standardize the **routine objective documentation** of motor recovery, 2) facilitate **communication** between and within **clinical** and **research** settings, and 3) provide a useful **outcome measure** for evaluating the impact of interventions on motor recovery. By keeping the scoring simple, and the instrument as concise as possible, and by involving many therapists in the various stages of development and testing, we hope to develop an instrument that will be useful to clinicians & that will be incorporated into routine clinical practice to improve the efficiency of our information gathering. By carefully developing the content of STREAM, and by testing the reliability, validity and responsiveness of the measure, we hope to develop a tool specifically suited for evaluating the efficacy of therapeutic interventions in terms of their influences on the recovery of voluntary movement and basic mobility.

APPENDIX 3.2.4
Scoring Questionnaire

Please consider these four scoring options and indicate (below) the scoring that you prefer.

OPTION A

I. VOLUNTARY MOVEMENT OF THE LIMBS
- 0 unable to perform the test movement through any visible range (includes flicker)
- 1 able to perform partial range of movement, but requires assistance, stabilization or support to complete or performs full movement with marked deviation from normal pattern
- 2 completes the movement in a manner that is comparable to the unaffected side
- *X activity not tested (specify why)*

II. BASIC MOBILITY
- 0 unable to perform the test movement
- 1 able to perform part of the movement actively, but requires supervision or assistance from an aid or an individual to complete or performs any or all of the movement with marked deviation from normal
- 2 completes the movement independently and safely
- *X activity not tested (specify why)*

OPTION B (4 point scale)

I. VOLUNTARY MOVEMENT OF THE LIMBS
- 0 unable to perform the test movement through any visible range (includes flicker)
- 1 able to perform partial range of movement, but requires assistance, stabilization or support to complete
- 2 performs full movement, but with marked deviation from normal pattern
- 3 completes the movement in a manner that is comparable to the unaffected side
- *X activity not tested (specify why)*

II. BASIC MOBILITY
- 0 unable to perform the test movement
- 1 able to perform part of the movement actively, but requires supervision or assistance from an aid or an individual to complete
- 2 performs any or all of the movement, but with marked deviation from normal pattern
- 3 completes the movement independently and safely
- *X activity not tested (specify why)*

OPTION C (reversal of caregories 1 & 2)

I. VOLUNTARY MOVEMENT OF THE LIMBS
- 0 unable to perform the test movement through any visible range (includes flicker)
- 1 performs full movement, but with marked deviation from normal pattern
- 2 able to perform partial range of movement with normal pattern, but requires assistance, stabilization or support to complete
- 3 completes the movement in a manner that is comparable to the unaffected side
- *X activity not tested (specify why)*

II. BASIC MOBILITY
- 0 unable to perform the test movement
- 1 performs any or all of the movement, but with marked deviation from normal pattern
- 2 able to perform part of the movement actively, but requires supervision or assistance from an aid or an individual to complete
- 3 completes the movement independently and safely
- *X activity not tested (specify why)*

OPTION D (any other suitable scoring scheme you would recommend or suggest)

_____

_____

_____

_____

_____

_____

Please indicate the scoring option that you prefer (circle choice), and feel free to add any comments as to the reasons for your choice:

A	B	C	D

✓

APPENDIX 3.2.5
Rating Form

Please rate each of the suggested modifications listed as to importance using the following key:

1 essential; **2** very important; **3** moderately important; **4** neither important nor unimportant; **5** **not** necessary.

Suggestion #	Rating
1	1 2 3 4 5
2	1 2 3 4 5
3	1 2 3 4 5
4	1 2 3 4 5
5	1 2 3 4 5
6	1 2 3 4 5
7	1 2 3 4 5
8	1 2 3 4 5
9	1 2 3 4 5
10	1 2 3 4 5
11	1 2 3 4 5
12	1 2 3 4 5
13	1 2 3 4 5
14	1 2 3 4 5
15	1 2 3 4 5
16	1 2 3 4 5
17	1 2 3 4 5
18	1 2 3 4 5
19	1 2 3 4 5
20	1 2 3 4 5
21	1 2 3 4 5
22	1 2 3 4 5
23	1 2 3 4 5
24	1 2 3 4 5
25	1 2 3 4 5
26	1 2 3 4 5
27	1 2 3 4 5
28	1 2 3 4 5
29	1 2 3 4 5
30	1 2 3 4 5
31	1 2 3 4 5
32	1 2 3 4 5
33	1 2 3 4 5
34	1 2 3 4 5
35	1 2 3 4 5
36	1 2 3 4 5
37	1 2 3 4 5
38	1 2 3 4 5
39	1 2 3 4 5
40	1 2 3 4 5

APPENDIX 3.2.6
Follow-Up Scoring Questionnaire

Thank you, once again, for your participation on this project of refining the STroke REhabilitation Assessment of Movement (STREAM)! We would appreciate additional input from you to confirm the appropriateness of the revised scoring of the STREAM.

Rationale For Revised Scoring: We wanted the scoring to be as simple as possible, in order to facilitate test administration and to optimize reliability. We wanted the scoring to reflect the return of voluntary movement; the level of independence and the general quality of movement should also be reflected. Also, keep in mind that the STREAM is intended to be used as an outcome measure (ie. concise, reliable...).

I. VOLUNTARY MOVEMENT OF THE LIMBS

Sample Items: raises hand to touch top of head; extends knee in sitting; dorsiflexes ankle in sitting...

Scoring:
- **0** **unable** to perform the test movement through any visible range (includes flicker)
- **1** **a.** able to perform only **partial** range of movement (requires assistance or stabilization to complete)
  b. performs **full** movement, but with **marked deviation** from normal pattern
  c. both a & b apply
- **2** **completes** the movement in a manner that is **comparable to the unaffected side**
- ***X*** *activity not tested (specify why)*

II. BASIC MOBILITY

Sample Items: rises to standing from sitting; takes 3 steps forwards; walks 10 meters....

Scoring:
- **0** **unable** to perform the test activity through any visible range (minimal active participation)
- **1** **a. able to perform only part of** the activity
  b. able to perform any or all of the activity, but with **marked deviation from normal**
  c. both a & b apply
- **2** **requires an aid to complete** the activity **independently and safely** (including orthotics)
- **3** **completes** the activity **independently and safely with no aid**
- ***X*** *activity not tested (specify why)*

Please select (circle) the statement which best describes how you feel about the revised scoring:

1) **acceptable as is**
2) **acceptable with minor changes *such as:***_____
3) **requires major changes** *such as:*_____

Please fax this form to me at the School of P&OT (**FAX # 398-8193**), as soon as is conveniently possible for you (we are awaiting your reply before going ahead with the reliability study). If you have any questions, or require further information, please contact **Kathy Daley @ 286-0844**.

Thanks sincerely, Kathy Daley

APPENDIX 3.3

Characteristics of Concensus Panel Participants (N =20)

Panel	Service Area	Caseload	# of Stroke Patients Assessed in Past Year	Years of Experience Physio (Stroke)
I	Acute	Mixed	< 10	12.5 (7.5)
I	Acute	Mixed	10-30	18 (3)
II	Acute	Mixed	< 10	17 (17)
II	Acute	Neuro	10-30	12 (12)
II	Acute	Neuro	10-30	4.5 (2.5)
I	In-pt rehab	Stroke	10-30	8 (3)
I	In-pt rehab	Neuro	< 10	5.5 (4)
I	In-pt rehab	Stroke	10-30	9 (3.5)
I	In-pt rehab	Neuro	> 30	28 (20)
I	In-pt rehab	Mixed	< 10	13 (10)
I	In-pt rehab	Mixed	10-30	16 (16)
I	In-pt rehab	Neuro	> 30	20 (10)
II	In-pt rehab	Stroke	10-30	7 (6)
II	In-pt rehab	Mixed	10-30	2 (2)
II	In-pt rehab	Stroke	> 30	4 (2)
I	Out-pt rehab	Neuro	> 30	32 (17)
I	Out-pt rehab	Mixed	10-30	5.5 (3)
II	Out-pt rehab	Neuro	> 30	24 (10)
II	Day Hospital	Mixed	10-30	5.5 (3.5)
II	LTC	Neuro	> 30	10 (4)

APPENDIX 4

Panels One and Two's Suggestions; Decisions Made and Rationale

TABLE OF CHANGES SUGGESTED BY FIRST CONCENSUS PANEL

# Rating As Important (N=11)	ITEM UNDER CONSIDERATION	DECISION & RATIONALE
10	Functional U/E task (eg. reach & grasp)	–; influenced by neglect etc. (would use specific/detailed U/E assessment)
10	*Hand items (graded: partial / full)*	+; improved sensitivity
10	*Walking items (graded: 'step to'/ 'step through')*	+; scoring changed (include 'qualitative'/marked deviation)
9	Lateral side flexion in sitting (trunk control)	–; trunk control covered in item 'reaching to the floor
9	Maintain knee position in 1/2 crooklying	–; covered in 'bending hip and knee in supine', and 'bridging'
9	Bridging	+; fundamental for bed mobility
9	Holding arm vertical in supine	–; difficult to standardize (ie. stabilization of elbow, facilitation, etc)
9	Moving from lying to sitting	+; fundamental for functional mobility / trunk control
9	*Oppose thumb to index finger*	+; more specific than opposition to all four fingers (graded difficulty)
7	*Oppose thumb to little finger*	+; more specific than opposition to all four fingers (graded difficulty)
8	Reach forward-sideways in sitting	–; difficult to standardize, sitting balance covered in 'reaching to the floor'
6	Roll to unaffected side	+; fundamental for functional mobility / trunk control
6	Roll to affected side	+; fundamental for functional mobility / trunk control
5	Raise arm in supine	–; more functional in sitting
5	Longer walk (50" / 15m)	–; endurance may be limiting factor (would use timed walk instead)
5	Lateral reach (sitting balance)	–; difficult to standardize; sitting balance covered in 'reaching to floor'
4	Crook-lying knees side-to-side	–; covered in 'rolling' and 'moving supine to sit'
3	Tandem standing	–; not functional; would use specific balance assessment
3	*Hand to ipsilateral ear*	+; modified to 'touch forehead' instead of ear (higher level)
3	*Turn 90 degrees in standing*	+; functional activity
2	*Sitting unsupported (change to > 1 minute)*	–; if can sit 60 seconds can probably sit longer
1	Abduction of unaffected leg	–; covered in functional mobility item 'stepping sideways to unaffected side'

DECISION KEY: + = ADDED OR RETAINED; – = DELETED, NOT ADDED OR UNCHANGED

ITALICS = EXISTING ITEM BEING RECONSIDERED

TABLE OF CHANGES SUGGESTED BY SECOND CONCENSUS PANEL

# Rating As Important *(N=9)*	ITEM UNDER CONSIDERATION	DECISION & RATIONALE
9	Sit unsupported (30 seconds)	+; fundamental for mobility in sitting
7	Stand (30 seconds)	+; fundamental for mobility in standing
7	Reach with unaffected hand to affected foot	--; similar to item reaching to unaffected foot
6	Hold head up in sitting	--; extremely low level
6	Keep hand on lap	--, influenced by neglect, sensation, clothing, etc.
6	*Knee flexion in standing*	+; item retained (unique item. selective movement)
6	*Supine to sit to both sides (two items)*	--; to either side acceptable (minimizes effect of neglect)
6	Maintain half crooklying	+, low level activity for L/E (hip rotary control)
5	*Rolling (clarify pattern)*	+, not pulling with hands (using trunk versus upper extremities)
---	---	---
4	Place hand on knee (from plinth)	--, could involve compensatory scapular elevation / trunk side bend
4	Pick up pen and place it poised to write	--, involves cognition, perception, sensation, etc.
4	Place hand on opposite shoulder	--; similar to item 'hand to top of head'
3	Trunk rotation (supine)	-, covered in supine mobility items
3	Trunk rotation (sitting)	--, covered in sitting mobility items
3	*Stairs (separate 'alternating'/'non alternating')*	, would result in excessive weight for stair climbing
3	Standing to sitting	, covered in sit to stand
3	*Sit to stand (without using hands or armrests)*	+, must not push off for full marks (high level activity)
1	*Wrist extension with finger extension*	--, too much weight on high level U/E activity, limited by ROM
1	*Oppose thumb to little finger*	--, too much weight on high level U/E activity, not functional
1	Neck rotation in supine	-, too low level, limited by ROM
1	Lift head up in supine	--, too low level
0	Trunk rotation (standing)	-, covered in standing mobility items, difficult to standardize
0	*Hip abduction in standing (affected side)*	+, limited number of L/E items, unique / selective / high level

DECISION KEY: + = ADDED OR RETAINED; – = DELETED, NOT ADDED OR UNCHANGED
ITALICS = EXISTING ITEM BEING RECONSIDERED

APPENDIX 5.1
Consent Forms for Participation in The Direct Observation Reliability Study

Testing the Measurement Properties of the Stroke Rehabilitation Assessment of Movement (STREAM)

Researchers at the Jewish Rehabilitation Hospital, in collaboration with researchers at McGill University, are conducting a study on how people regain the ability to move the arm and leg after a stroke. We are asking you to participate in this study. If you agree to participate, we will ask you to let a specially trained physical therapist evaluate how well you can move your limbs and carry out other activities like standing and walking. At the same time, another therapist will observe. The evaluation will take less than one half hour of your time and will be done outside of your regular therapy time, so that there will be no interruption of treatment time.

All information collected in this study is strictly confidential. This means that your name will never be identified in any publications or presentations of the findings of this research. Whether you agree to participate or not will not affect your treatment or any other aspect of your stay here. In addition, if you agree to participate and later change your mind, you may withdraw from the study without any negative consequences to you. We would also like to make clear to you that your participation in this study is to benefit research and, while in the future it may help other patients with strokes, it will not benefit you directly. Any questions or concerns you may have about this study may be directed to Kathy Daley (phone 286-0844), or to Martha Visintin through the research department at the Jewish Rehabilitation Hospital (phone 688-9550).

I understand what is required of me, and I agree to participate in this study.

Patient's signature:_____

Witness:_____

Date_____

FORMULAIRE DE CONSENTEMENT

STREAM

Les chercheurs de l'Hôpital Juif de Réadaptation, en collaboration avec des chercheurs de l'Université McGill effectuent actuellement une recherche portant sur la récupération de la mobilité du bras et de la jambe suite à un accident cérébro-vasculaire.

Nous sollicitons votre participation à cette étude. Si vous acceptez d'y participer, nous vous demandons de permettre à un thérapeute qualifié d'évaluer la façon dont vous faites certains mouvements et certaines activités, telles que la marche et autres activités en position debout. Un second thérapeute sera présent lors de l'évaluation seulement puor observer. Cette évaluation prendra moins de 30 minutes de votre temps et se fera en dehors de votre temps régulier de thérapie, afin que vos heures de traitements ne soient pas interrompues.

Toutes les informations recueillies au cours de cette étude seront strictement confidentielles. Votre nom n'apparaîtra jamais dans aucune publication ou présentation portant sur les résultats de cette étude.

Votre participation à cette étude est volontaire et vous pouvez en tout temps vous retirer sans que votre retrait affecte votre traitement à l'hôpital. Nous désirons également souligner que vous ne bénéficierez pas directement de cette étude mais que votre participation permettra dans l'avenir aider d'autres personnes ayant subi un accident cérébro-vasculaire.

Si vous désirez obtenir de plus amples détails ou des informations additionnelles concernant cette étude, n'hésitez pas à communiquer avec Kathy Daley (286-0844), ou Martha Visintin à l'Hôpital Juif de Réadaptaion (688-9550).

Votre signature apposée ci-dessous indique que vous avez lu, ou qu'on vous a lu ce document, que vous le comprenez et que vous consentez à participer à cette étude.

Signature du patient:_____

Représentant des chercheurs:_____
(témoin)

Date_____

APPENDIX 5.2
Authorization for Photographs, Films, Audio Cassettes, and Videotapes

I, _____

Authorize the hospital _____

To make:	yes	no
Photographs	___	___
Films	___	___
Audio Cassettes	___	___
Videotapes	___	___
Other (specify)_____	___	___

and I authorize the researcher to use or publish these films for scientific or educational purposes, with the condition that all reasonable precautions will be taken to preserve my anonymity.

I however put forward the following restrictions (if any reservations):

_____

_____

_____

_____ _____

Signature of patient Date

_____ _____

Witness Date

Autorisation Pour Phothographies, Films, Enregistrements Sonores, Enregistrements Sur Bande Magnétoscopique et Autres

Je, soussigé_____

Autorise le centre hospitalier_____

À faire les:	Oui	Non
Photographies	___	___
Films	___	___
Enregistrements sonores	___	___
Enregistrements sur bande magnétoscopique	___	___
Autres (préciser)_____	___	___

et je autorise les chercheurs à utiliser et à publier ces films médicales, scientifiques et éducatives, à la condition que des précautions raisonnables soient prises pour que soit conservé l'anonymat.

J'émets cependant les restrictions suivantes:

_____

_____

_____

_____ _____

Signature du patient Date

_____ _____

Témoin Date

APPENDIX 6

Documents from The Reliability Studies

APPENDIX 6.1
Scoring Form: test version of STREAM

Score	
	SUPINE
	1. PROTRACTS SCAPULA IN SUPINE
/2	*"Lift your shoulder blade so that your hand moves towards the ceiling"*
	Note: therapist stabilizes arm with shoulder 90° flexed and elbow extended.
	2. EXTENDS ELBOW IN SUPINE (starting with elbow fully flexed)
/2	*"Lift your hand towards the ceiling, straightening your elbow as much as you can"*
	Note: therapist stabilizes arm with shoulder 90° flexed; strong associated shoulder extension and/or abduction = marked deviation (score 1a or 1c).
	3. MAINTAINS AFFECTED LEG IN HALF CROOK LYING POSITION FOR 10 COUNTS
	"Keep your leg in this position while I count to ten"
/2	Note: therapist places patient in half crook lying (ie. hip and knee flexed, with foot flat on support surface, with hip in neutral rotation) and holds foot in place; if unable to maintain the knee steady for a count of ten = marked deviation (score 1a if able to partially correct knee position, but not to midline; 1c if able to repeatedly correct to midline position).
	4. FLEXES HIP AND KNEE IN SUPINE (ATTAINS HALF CROOK LYING)
/2	*"Bend your hip and knee so that your foot rests flat on the bed"*
	5. ROLLS ONTO SIDE (starting from supine)
/3	*"Roll onto your side"*
	Note: may roll onto <u>either</u> side; pulling with arms to turn over = aid (score 2)
	SIDE LYING (on unaffected side)
	6. ABDUCTS AFFECTED HIP IN SIDE LYING
/2	*"Lift your foot up to here (therapist holds hand approximately 20 cm above foot), keeping your hip & knee straight and your pelvis perpendicular to the bed"*
	Note: unaffected leg is partially flexed to assist balance.
	CROOK LYING
	7. MOVES KNEES SIDE TO SIDE IN CROOK LYING (TRUNK ROTATION)
/3	*"Keep your knees together and lower them as far as you can over to one side, then over to the other side, and return them to the middle; keep your shoulders resting on the bed"*
	Note: active movement to <u>both</u> sides must be full (comparable to passive rotation) to obtain full marks; if requires aid (external or from therapist) to maintain knees together = aid (score 2)
	8. RAISES HIPS OFF BED IN CROOK LYING (BRIDGING)
	"Lift your hips as high as you can"
/3	Note: therapist may stabilize foot, but if knee pushes strongly into extention with bridging = marked deviation (score 1a or 1c); if requires aid (external or from therapist) to maintain knees in midline = aid (score 2)
	9. MOVES FROM LYING SUPINE TO SITTING WITH FEET ON THE FLOOR
/3	*"Sit up and place your feet on the floor"*
	Note: may sit up to <u>either</u> side using any functional and safe method; longer than 20 seconds = marked deviation (score 1a or 1c); pulling up using bedrail or edge of plinth = aid (score 2)

SITTING (feet supported; hands resting on pillow on lap for items 10-21)

/3 10. MAINTAINS ERECT SITTING FOR 20 COUNTS
"Sit as straight as you can while I count to twenty"
Note: holding on with hand(s) = aid (score 2)

/2 11. SHRUGS SHOULDERS (SCAPULAR ELEVATION)
"Shrug your shoulders as high as you can"
Note: both shoulders are shrugged simultaneously.

/2 12. RAISES HAND TO TOUCH TOP OF HEAD
"Raise your hand to touch the top of your head"

/2 13. PLACES HAND ON SACRUM
"Reach behind your back and as far across toward the other side as you can"

/2 14. RAISES ARM FORWARD TO 90° OF FORWARD FLEXION
"Reach your arm straight out in front of you to horizontal"

/2 15. RAISES ARM OVERHEAD TO FULLEST ELEVATION
"Reach your hand as high as you can towards the ceiling"

/2 16. RAISES ARM SIDEWAYS TO 90° OF ABDUCTION
"Lift your arm sideways to horizontal, keeping your elbow straight"

/2 17. SUPINATES AND PRONATES FOREARM WITH ELBOW FLEXED AT 90°
"Keeping your elbow bent and close to your side, turn your forearm over so that your palm faces up, then turn your forearm over so that your palm faces down"

/2 18. CLOSES HAND FROM FULLY OPENED POSITION
"Make a fist, keeping your thumb on the outside"
Note: must extend wrist slightly (ie. power grip with wrist cocked) to obtain full marks.

/2 19. OPENS HAND FROM FULLY CLOSED POSITION
"Now open your hand all the way"

/2 20. OPPOSES THUMB TO INDEX FINGER (tip to tip)
"Make a circle with your thumb and index finger"

/2 21. EXTENDS WRIST AND FINGERS
"Keeping your forearm resting on the pillow, lift your hand and straighten your fingers"

/2 22. FLEXES HIP IN SITTING
"Lift your knee as high as you can"

/2 23. EXTENDS KNEE IN SITTING
"Straighten your knee by lifting your foot up"

/2	24. FLEXES KNEE IN SITTING *"Slide your foot back under you as far as you can"* **Note:** start with affected foot forward (heel in line with toes of other foot).
/2	25. DORSIFLEXES ANKLE IN SITTING *"Keep your heel on the ground and lift your toes off the floor as far as you can"*
/2	26. PLANTARFLEXES ANKLE IN SITTING *"Keep your toes on the ground and lift your heel off the floor as far as you can"*
/2	27. EXTENDS KNEE AND DORSIFLEXES ANKLE IN SITTING *"Straighten your knee and bring your toes towards you"*
/3	28. REACHES FORWARD TO TOUCH UNAFFECTED FOOT WITH UNAFFECTED HAND *"With your strong hand, reach down and touch your foot on the same side"* **Note:** pushing up with hand(s) to return to sitting = aid (score 2)
/3	29. RISES TO STANDING FROM SITTING *"Stand up; try to take equal weight on both legs"* **Note:** pushing up with hand(s) to stand = aid (score 2)
/3	STANDING 30. MAINTAINS STANDING FOR 20 COUNTS *"Stand on the spot while I count to twenty"*
/2	**STANDING (holding onto a stable support to assist balance for items 31-33)** 31. ABDUCTS AFFECTED HIP WITH KNEE EXTENDED IN STANDING *"Keep your knee straight and your hips level, and raise your leg to the side"*
/2	32. FLEXES AFFECTED KNEE WITH HIP EXTENDED IN STANDING *"Keep your hip straight and bend your knee back; bring your heel towards your bottom"*
/2	33. DORSIFLEXES AFFECTED ANKLE IN STANDING WITH KNEE EXTENDED *"Keep your heel on the ground and lift your toes off the floor as far as you can"*
/3	**STANDING AND WALKING ACTIVITIES** 34. PLACES AFFECTED FOOT ONTO FIRST STEP (or onto stool approximately 18 cm high) *"Lift your foot and place it onto the first step (or stool) in front of you"* **Note:** returning the foot to the ground is not scored; use of handrail = aid (score 2)
/3	35. TAKES 3 STEPS FORWARDS (one and a half gait cycles) *"Take three average sized steps forwards, placing one foot in front of the other"*
/3	36. TAKES 3 STEPS BACKWARDS (one and a half gait cycles) *"Take three average sized steps backwards, placing one foot behind the other"*

/3	37. TAKES 3 STEPS SIDEWAYS TO UNAFFECTED SIDE *"Take three average sized steps sideways towards your strong side"*
/3	38. TAKES 3 STEPS SIDEWAYS TO AFFECTED SIDE *"Take three average sized steps sideways towards your weak side"*
/3	39. MOVES FEET TO TURN BODY 90° TOWARDS UNAFFECTED SIDE *"Take steps to turn 90° on the spot towards your strong side"*
/3	40. MOVES FEET TO TURN BODY 90° TOWARDS AFFECTED SIDE *"Take steps to turn 90° on the spot towards your weak side"*
/3	41. WALKS 10 METERS INDOORS (on smooth, obstacle free surface; within 20 seconds) *"Walk in a straight line over to ... (a specified point 10 meters away) "* **Note:** orthotic=aid (score 2); longer than 20 seconds =marked deviation (score 1c)
/3	42. WALKS UP 3 STAIRS ALTERNATING FEET *"Walk up three stairs; place only one foot at a time on each step if you can"* **Note:** handrail=aid (score 2); non-alternating feet = marked deviation (score 1a or 1c)
/3	43. WALKS DOWN 3 STAIRS ALTERNATING FEET *"Walk down three stairs; place only one foot at a time on each step if you can"* **Note:** handrail=aid (score 2); non-alternating feet = marked deviation (score 1a or 1c)

APPENDIX 6.2
Therapist Profile

PLEASE PROVIDE US WITH THE FOLLOWING INFORMATION:

Name (will remain confidential): _____

Phone # (H)

(W)

Place of Employment: _____

Current Service Area:

i) Staff Therapist___ Senior___ Clinical Specialist___ Other_____

ii) Acute Care___ In-patient Rehab.___ Out-patient Rehab.___ LTC___

Other (specify) _____

iii) General (Mixed Caseload)___ Mixed Neuro ___ Stroke Service___

Other_____

Number of Stroke Patients Assessed in the Past Year: < 10 ___ 10-30___ > 30___

Years of Experience (to closest 1/2 year):

In current service area _____

In stroke related service _____

Total (as a physio) _____

Training: BScPT_____ Other (specify)_____

University where Degree Obtained: _____

APPENDIX 6.3

INSTRUCTIONS FOR RATERS

THANK YOU! for your participation in this study to evaluate the reliability of the STroke REhabilitation Assessment of Movement (STREAM). You will be asked to rate 4 subjects using the assessment forms provided. Your individual ratings will remain confidential.

Please rate each patient on the sheet provided (the subjects' number will be given at the beginning of each taped session; each sheet is labelled with a corresponding subject number). Please score each item for each patient according to your understanding of the directions for scoring, and based on what you see the patient perform on the tape.

You will have only one opportunity to view each taped session. No discussion or explanations will be given while the tapes are being viewed, as we want to assess how the STREAM items are interpreted by each individual. In the videotapes the items have been assessed in the same order as presented on the form. There will be a short break after viewing the first two subjects.

THANKS AGAIN for your valued expertise!

APPENDIX 7

Characteristics of Raters Participating in the Direct Observation Reliability Study

Service Area	Caseload	# of Stroke Patients Assessed in Past Year	Years of Experience Physio (Stroke)
Acute	Neuro	10-30	4.5 (2.5)
In-pt rehab	Stroke	>30	4.5 (2.5)
In-pt rehab	Mixed	>30	2 (2.5)
In-pt rehab	Neuro	>30	4.5 (1.5)
In-pt rehab	Brain injury	<10	9 (3)
Mixed (grad student)	Neuro	10-30	3.5 (3.5)

APPENDIX 8

Test Manual

STroke REhabilitation Assessment of Movement

(STREAM)

School of Physical and Occupational Therapy
Davis House, McGill University
3654 Drummond, Montreal, PQ,
H3G 1Y5, Canada

Physiotherapy Department
Jewish Rehabilitation Hospital
3205 Alton Goldbloom, Chomedy, PQ,
H7V 1R2, Canada

APPENDIX 8.1
Background Information

BACKGROUND OF STREAM

The original STREAM was developed in 1986 through a collaborative effort between researchers and physical and occupational therapists at the Jewish Rehabilitation Hospital (JRH), Chomedy-Laval, Quebec. The items included in this original instrument were selected from clinical experience and from existing published assessments (Ashburn, 1982; Bobath, 1978; Brunnstrom, 1970; Fugl-Meyer, 1975). The STREAM has since been revised through studies conducted by researchers at McGill University and the JRH. Content validity, and preliminary reliability of STREAM, have been established, and clinical use of the instrument is supported.

INTENDED PURPOSE & SCOPE

The purpose of the STREAM is to provide a relatively comprehensive, standardized, objective and quantitative clinical measure of the motor function of stroke patients. It is designed to be used by physiotherapists for **monitoring motor recovery**, and for **evaluating the impact of therapies,** such as physical therapy, medical and pharmacologic interventions, etc. on the motor recovery of stroke patients. The STREAM is intended for use as an **outcome measure** for motor recovery-- it is not intended for use as an assessment tool for diagnosis or treatment planning, if used in isolation. STREAM may, however, be used in conjunction with measures of spasticity, ROM, sensation, balance, functional independence in ADL, etc. to obtain a more global profile of a patient's physical function.

The STREAM is designed to be suitable for evaluating stroke patients with motor deficits ranging from mild to severe. The test items represent gradations of difficulty, to allow discrimination between patients with differing levels of motor impairment, and to monitor an individual's motor recovery over time. The items included in the STREAM cover a wide range of motor activities that are typically assessed by physiotherapists for the purpose of evaluating the motor status of stroke patients, specifically **voluntary movement of the limbs** and **basic mobility**. The items assessing limb movement patterns are intended to provide a profile of basic motor status (motor recovery in the limbs); the remaining items give a more functional picture of motor recovery, assessing simple motor tasks which require the integrated movement of trunk and limbs.

Since the focus of STREAM is the assessment of movement recovery (specifically motor ability), the items included in the STREAM have been limited to simple movements and activities. In an attempt to minimize the confounding effects that variables other than motor recovery, such as cognition,

perception and communication, have on the measurement of motor status, some of the mobility items (rolling and moving from supine to sitting) may be performed to either side. Similarly, only simple one or two step commands are used, and may be supplemented by demonstration to facilitate communication.

The STREAM has intentionally been designed such that a minimum of equipment is utilized, in order to improve the portability and clinical utility of the instrument. STREAM is also brief (10-15 minutes) and easy to administer.

SCORING

The scoring of STREAM is designed to be objective and quantitative, so that a high reliability or reproducibility of scores should be achievable. A number of scoring schemes or dimensions were incorporated, including scoring with respect to: the degree of assistance required or level of independence, active range of voluntary movement possible, and quality of movement. We wanted the scoring to be as simple and unambiguous as possible so that, after reading the instruction booklet, therapists would be able to carry out the testing accurately and easily using only a short scoring form. Space is provided on the scoring form for documenting a patient's STREAM scores on four occasions.

For functional purposes, it is essential that a patient is able to perform an activity independently and safely. In some cases, however, variables such as cognition and perception may play a greater role than motor recovery in a patient's functional ability. Thus, on STREAM, a patient will be given full credit for their performance, even if they require supervision due to cognitive or perceptual deficits, if they have the motor ability to achieve a given task. To test a patient's functional independence, where the contributions of motor, cognitive, and perceptual functioning are considered simultaneously, other more suitable measures should be selected.

The STREAM includes a total of 30 items, with 20 items assessing limb movement patterns (10 for the lower extremity and 10 for the upper extremity) and 10 items assessing basic mobility. The STREAM's scoring scheme, while remaining simple to preserve reproducibility, provides information as to both the amount and quality of movement. A three point ordinal scale is employed for scoring voluntary movement of the limbs; a patient is scored **0** if unable to perform the test movement, **1** if a movement can be only partially completed or if the movement is performed with marked deviation from a normal pattern when compared to the unaffected limb, and **2** if able to complete the movement in a manner that is qualitatively and quantitatively near normal. Scoring category **1** is divided into three sub-categories: **1a** for movements performed only partially and in an abnormal pattern, **1b** for movements performed only partially but with a normal pattern, and **1c** for movements performed completely but with an abnormal pattern. The same scoring scheme is employed for the basic

mobility subscale, except that a category has been added to allow for independence with the help of an aid. If a test item cannot be performed due to pain or limited passive range, or if a movement or activity is limited because of other reasons such as perceptual or cognitive deficits, amputation, impaired vision, etc. then this must be indicated by scoring an X and indicating the reason. Details are provided in the specific directives for scoring STREAM.

CONCEPTUAL FRAMEWORK

In devising a clinical instrument to measure motor recovery, a number of aspects have to be considered, such as: the domains to be included, the methods by which these are assessed, and the uses and populations for which the measure is appropriate. First, motor recovery must be conceptually and operationally defined. However, given the complexity of motor performance in terms of the number of factors that potentially influence motor recovery and functional performance following strokes, defining and measuring this phenomenon presented a challenge. A variety of terms could have been used to describe what the STREAM endeavors to measure, including: motor recovery, motor ability, motor function, functional motor recovery, motor performance, motor control, motricity, and motor output. Unfortunately, many of these terms are not well defined and may imply different concepts to different individuals. We selected the terms "voluntary movement and basic mobility" to describe the attributes measured by the STREAM, and the following conceptual framework was developed.

Motor recovery implicates a number of neurophysiologic processes at the cellular level. In addition, factors such as the size and side of the lesion, age, comorbid conditions, motivation, communication, cognition, contractures and pain impact on motor recovery. Although it is very difficult to separate out the relative influences of these various processes and factors, motor recovery is manifested by the re-emergence of voluntary movement and restoration of basic mobility. Therefore, the STREAM is intended to measure these fundamental building blocks that reflect motor recovery, and that can be easily measured in the clinical setting.

In the context of physical rehabilitation following stroke, impairments refer to the primary deficits observed such as paresis, sensory deficits and abnormal tone which contribute to, or cause, motor dysfunction, whereas disability refers to the functional consequences of the impairment such as reduced ability to perform purposeful movements, transfers, walking, and activities of daily living (World Health Organization, 1980). In this framework, the STREAM is related hierarchically to other measures of impairments and disabilities in that it is intended to measures basic motor ability - one step beyond the level of the primary impairments, and one step before functional mobility and ADL measures.

EXPECTED CONTRIBUTIONS

The STREAM should: 1) streamline and standardize the **routine objective documentation** of motor recovery, 2) facilitate **communication** between and within **clinical** and **research** settings, and 3) provide a useful **outcome measure** for evaluating the impact of interventions on motor recovery. By keeping the scoring simple, the instrument concise, and by involving many therapists in development and testing, we hope we have produced an instrument that will be useful to clinicians, and that will be incorporated into routine clinical practice to improve the efficiency of information gathering. We also hope that, by carefully developing the content of STREAM, and, following further testing of the reliability, validity and responsiveness of the measure, STREAM will be a useful research tool specifically suited for evaluating the efficacy of therapeutic interventions in terms of their influences on the recovery of voluntary movement and basic mobility following stroke.

STAGES OF DEVELOPMENT

The process of scale development is an ongoing enterprise. The first stage of this project, involving content validation, preliminary reliability testing, and internal consistency analysis is now complete. As a first step in this project, we carried out a content verification survey to assess the broad acceptability of the STREAM. As a second step, the STREAM was presented to several panels of physiotherapists, who revised the instrument based on their collective clinical experience. Inter-rater reliability was assessed by comparing scores on STREAM made by direct observation of patients in the clinical setting, as well as through use of videotaped assessments of patients. Intra-rater reliability was also assessed using videotaped patient assessments, viewed on two occasions. Reliability was excellent both within and between raters; generalizability correlation coefficients (GCCs) for total STREAM scores were 0.99, and ranged from 0.963--0.998 for the subscale scores, for both the direct observation and videotaped assessments reliability studies. For individual items, Kappa statistics clustered in the range from 0.8 to 1.0, with only one of the 30 items demonstrating less than excellent agreement. Internal consistency was excellent, with Cronbach's Alphas of greater than 0.98 on subscales and total scores. Stages to follow will include further attention to content and construct validity through statistical analysis of the principle components of STREAM, as well as longitudinal studies to assess the responsiveness of the instrument.

For further information contact the STREAM Research Group:
Kathy Daley, Dr. Nancy Mayo, Dr Sharon Wood-Dauphinee, @ School of P&OT, McGill
Martha Visintin, Rosslyn Cabot @ JRH

APPENDIX 8.2

General Instructions for Using the STREAM

1. The patient should be in his/her usual state of attention and health.

2. The patient should be dressed in clothing that does not restrict movement, and that allows the therapist to observe the movement clearly (eg. shorts and T-shirt). Comfortable walking shoes or the patients usual footwear should be worn when testing the activities performed in standing.

3. Instructions (*italics* on scoring form) should be given verbally, demonstrated and repeated to the patient as necessary. For the items testing voluntary movement of the limbs, ask the patient to perform the movement once with the unaffected side. This allows you to observe the patient's comprehension of the test item, and the available range and movement pattern on the patient's 'unaffected side'.

4. If the patient's sitting balance is precarious, they may be seated on a chair with back support while testing items performed in sitting (items # 7-21).

5. Therapists may assist the patient to maintain standing while performing items # 23-25. Stabilization of the arm (items 1 & 2), and foot (item 5) is permitted where specified.

6. Therapists may assist the patient to achieve the starting positions specified. However, no hands-on *facilitation* of the limb movements should be given; if support or *partial* physical assistance (except as stated above in instructions 4 & 5) is required for performance of the mobility items, the patient is given a score of *1a or 1b.*

7. If necessary, the patient is permitted *three* attempts on each item and the best performance recorded.

8. The items should be tested in the order as presented.

9. Therapists should count at a rate such that 20 counts is equivalent to 20 seconds (eg. "one-1000, two-1000, three-1000....."; this should be timed and practiced several times prior to testing).

10. If the assessment is interrupted for any reason, it may be restarted from where it was left off if done so within a 24 hour period. If not, it should be redone from the beginning.

11. An item should be excluded (score X) if movement is limited by marked restriction of passive range or pain, and the following codes used to indicate the reason: **ROM, Pain, Other (reason).**

12. The following equipment should be available for use:
    - **sturdy stool** (or treatment plinth or armless chair) of a height such that patient can sit comfortably on a firm support with feet resting on the floor or on a small foot stool, with the hips and knees at 90°
    - **support surface** (firm, and large enough to permit rolling safely; raised approximately 1/2 meter off the ground); if using the patient's bed, it must be flat and encumbering bedding should be removed; alternatively, a large treatment plinth (raised mat) may be used
    - **pillow**
    - **stairs** with railings (departmental steps or full flight: standard height approximately 18 cm)

APPENDIX 8.3

STREAM SCORING

I. VOLUNTARY MOVEMENT OF THE LIMBS (/2)

0 **unable** to perform the test movement through any appreciable range (includes flicker or slight movement)

1 **a.** able to perform only **part** of the movement, and with **marked deviation** from normal pattern

b. able to perform only **part** of the movement, but in a manner that is **comparable to the unaffected side**

c. able to complete the movement, but only with **marked deviation** from normal pattern

2 able to complete the movement in a manner that is **comparable to the unaffected side**

X *activity not tested (specify why;* **ROM, Pain, Other (reason)***)*

II. BASIC MOBILITY (/3)

0 **unable** to perform the test activity through any appreciable range (ie. minimal active participation)

1 **a.** able to perform only **part** of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, and with **marked deviation** from normal pattern

b. able to perform only **part** of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, but with a grossly **normal** movement pattern

c. able to complete the activity independently, with or without an aid, but only with **marked deviation** from normal pattern

2 able to complete the activity independently with a grossly **normal** movement pattern, but **requires an aid**

3 able to complete the activity independently with a grossly **normal** movement pattern, **without an aid**

X *activity not tested (specify why;* **ROM, Pain, Other (reason)***)*

AMPLITUDE OF ACTIVE MOVEMENT

MOVEMENT QUALITY	None	Partial	Complete
Marked Deviation	0	1 a	1 c
Grossly Normal	0	1 b	2 (3)

Glossary of Terms for Scoring

(use as a reference the first few times you are scoring the STREAM)

Movement through an appreciable range: implies movement of an observable amplitude that is *greater than* a *flicker* or a small, essentially *nonfunctional*, movement (ie. must be at least 10%of the normal amplitude of movement).

Part of the movement (Limb movements: 1a & b): includes any active movement observed (without hands on facilitation) that is greater than a flicker or slight movement (category 0) and less than the complete movement (categories 1c, or 2).

Part of the activity (Basic mobility: 1a & b): implies that a patient is able to actively participate in a basic mobility activity (ie. does not require major assistance), but is unable to complete the activity without partial assistance or stabilization.

Complete movement (1c, 2 & 3): refers to movement that is *comparable* to the *quantity* of movement observed on the unaffected side, or to the attainment of a basic mobility task (ie. must be at least 90%of the normal amplitude of movement).

Marked deviation (1a & 1c): the performance of the test activity does not follow a natural sequence of movement comparable to how an individual without motor impairment would perform it (ie. it is not within the expected range of so called "normal movement"). Thus, moderate or major deviations or irregularities of movement, including strong associated reactions, gross postural asymmetry, and tremor or dysmetria *interfering with function*, should result in downgrading (ie. scores of 1a or 1c).

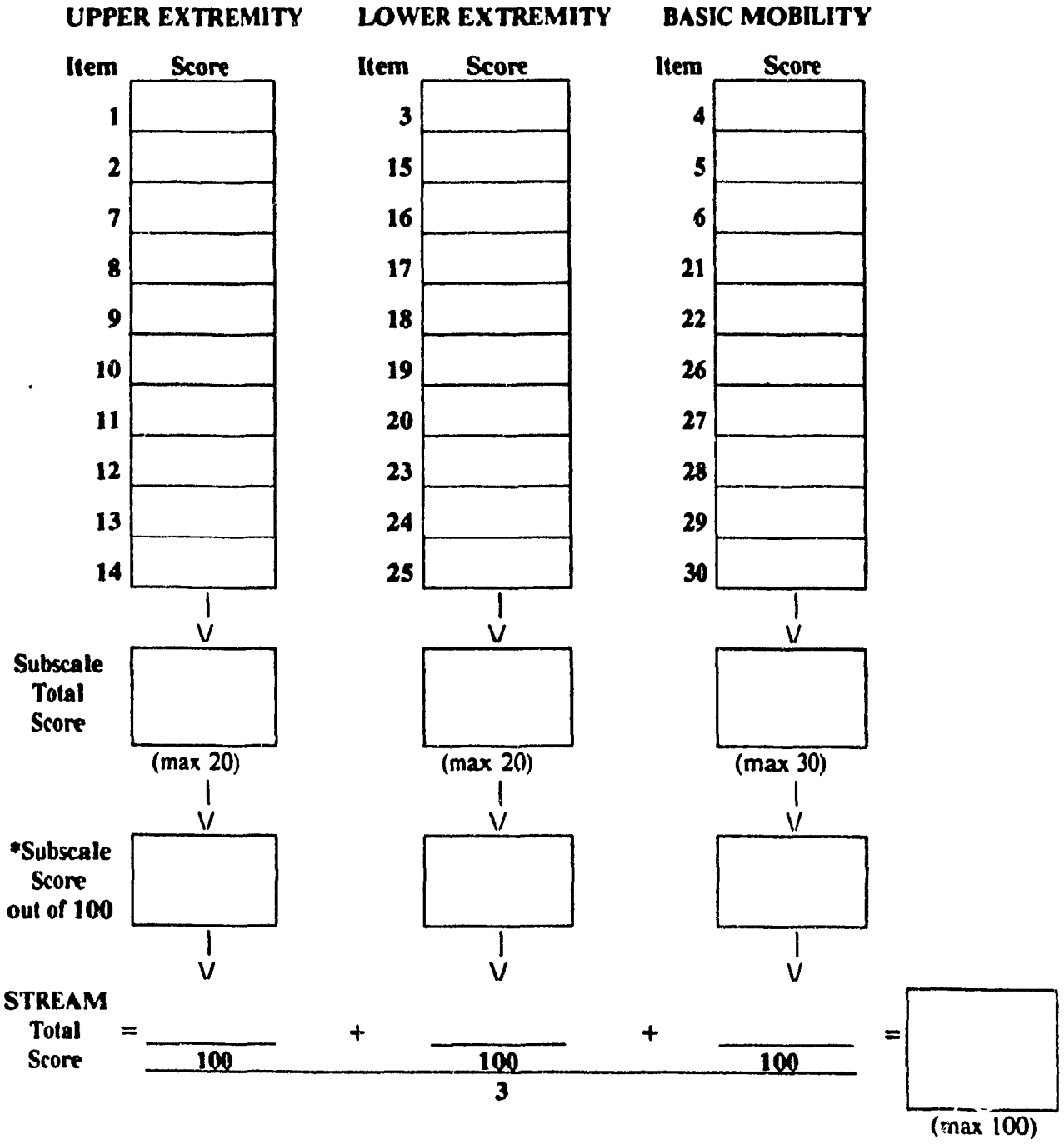
Comparable to the unaffected side (1b, 2 & 3): the performance of the test movement or activity *closely resembles* the *quality* and/or *quantity* of movement observed on the unaffected side.

Grossly normal movement pattern (1b, 2 & 3): the performance of the test activity follows a natural sequence of movement comparable to how an individual without motor impairment would perform it (ie. it is within the expected range of so called "normal movement"). Thus, to get full marks the movement need not be perfectly executed, but must be approaching normal; minor deviations or irregularities of movement should not result in downgrading.

Aid: refers to any *external / adaptive device(s)* (walking aids, splints, etc.) that may be used by a patient to perform a movement. The use of hand(s) to push up to stand, and the use of handrail(s) in stair climbing are also graded as using an aid.

Able to *complete* the activity *independently* (category II: 1c, 2 & 3): implies that a patient is able to carry out the basic mobility task without any *hands-on*, or *physical* assistance from another person; verbal encouragement, cueing and close supervision however may be given.

STREAM SCORE



Note: maximum score is based on the number of items scored. ie. for limb subscales, maximum score is 20 if all items are scored, 18 if only 9 items are scored, etc. Similarly, for mobility subscale, maximum score is 30, 27 if 9 items are scored, ...

*Subscale scores are transformed to a score out of 100 to correct for items not scored (due to pain, limited ROM, etc.). In addition, since the transformed subscale scores have the same denominator, equal weight is given to each of the subscales when the total STREAM score is obtained by summing the transformed subscale scores.

Example

STREAM SCORE

UPPER EXTREMITY		LOWER EXTREMITY		BASIC MOBILITY	
Item	Score	Item	Score	Item	Score
1	2	3	2	4	3
2	2	15	2	5	3
7	2	16	2	6	3
8	2	17	2	21	3
9	2	18	1c	22	3
10	X(Pain)	19	2	26	3
11	1	20	1a	27	3
12	2	23	2	28	3
13	2	24	1b	29	2
14	1b	25	X(Pain)	30	2
Subscale Total Score	16/18 (max 20)		15/18 (max 20)		28/30 (max 30)
*Subscale Score out of 100	89/100		83/100		93/100

STREAM Total Score = $\frac{\frac{89}{100} + \frac{83}{100} + \frac{93}{100}}{3} = \frac{88}{100}$ (max 100)

APPENDIX 8.4
The Thirty Item STREAM

STroke REhabilitation Assessment of Movement (STREAM)

Assessment Dates (Y/M/D)

1.
2.
3.
4.

Patient's Name:_____

Date of CVA:_____ **Sex: M F** **Age:**_____

Side of Lesion: L R **Side of Hemiplegia: L R**

Comorbid Conditions:_____

Type of aid(s) used:_____

Physiotherapist(s):_____

General Comments: _____

STREAM SCORING

I. VOLUNTARY MOVEMENT OF THE LIMBS

0 unable to perform the test movement through any appreciable range (includes flicker or slight movement)
1 a. able to perform only part of the movement, and with marked deviation from normal pattern
b. able to perform only part of the movement, but in a manner that is comparable to the unaffected side
c. able to complete the movement, but only with marked deviation from normal pattern
2 able to complete the movement in a manner that is comparable to the unaffected side
X *activity not tested (specify why;* ROM, Pain, Other (reason))

II. BASIC MOBILITY

0 unable to perform the test activity through any appreciable range (ie. minimal active participation)
1 a. able to perform only part of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, and with marked deviation from normal pattern
b. able to perform only part of the activity independently (requires partial assistance or stabilization to complete), with or without an aid, but with a grossly normal movement pattern
c. able to complete the activity independently, with or without an aid, but only with marked deviation from normal pattern
2 able to complete the activity independently with a grossly normal movement pattern, but requires an aid
3 able to complete the activity independently with a grossly normal movement pattern, without an aid
X *activity not tested (specify why;* ROM, Pain, Other (reason))

		AMPLITUDE OF ACTIVE MOVEMENT		
		None	Partial	Complete
MOVEMENT QUALITY	Marked Deviation	0	1 a	1 c
	Grossly Normal	0	1 b	2 (3)

SCORE					
4	3	2	1		
				/2	**SUPINE** 1. PROTRACTS SCAPULA IN SUPINE *"Lift your shoulder blade so that your hand moves towards the ceiling"* Note: therapist stabilizes arm with shoulder 90° flexed and elbow extended.
				/2	2. EXTENDS ELBOW IN SUPINE (starting with elbow fully flexed) *"Lift your hand towards the ceiling, straightening your elbow as much as you can"* Note: therapist stabilizes arm with shoulder 90° flexed; strong associated shoulder extension and/or abduction = marked deviation (score 1a or 1c).
				/2	3. FLEXES HIP AND KNEE IN SUPINE (attains half crook lying) *"Bend your hip and knee so that your foot rests flat on the bed"*
				/3	4. ROLLS ONTO SIDE (starting from supine) *"Roll onto your side"* Note: may roll onto either side, pulling with arms to turn over = aid (score 2).
				/3	5. RAISES HIPS OFF BED IN CROOK LYING (BRIDGING) *"Lift your hips as high as you can"* Note: therapist may stabilize foot, but if knee pushes strongly into extension with bridging = marked deviation (score 1a or 1c); if requires aid (external or from therapist) to maintain knees in midline = aid (score 2).
				/3	6. MOVES FROM LYING SUPINE TO SITTING (with feet on the floor) *"Sit up and place your feet on the floor"* Note: may sit up to either side using any functional and safe method; longer than 20 seconds = marked deviation (score 1a or 1c); pulling up using bedrail or edge of plinth = aid (score 2).
				/2	**SITTING** (feet supported; hands resting on pillow on lap for items 7-14) 7. SHRUGS SHOULDERS (SCAPULAR ELEVATION) *"Shrug your shoulders as high as you can"* Note: both shoulders are shrugged simultaneously.
				/2	8. RAISES HAND TO TOUCH TOP OF HEAD *"Raise your hand to touch the top of your head"*
				/2	9. PLACES HAND ON SACRUM *"Reach behind your back and as far across toward the other side as you can"*
				/2	10. RAISES ARM OVERHEAD TO FULLEST ELEVATION *"Reach your hand as high as you can towards the ceiling"*

SCORE 4	3	2	1		
				/2	11. SUPINATES AND PRONATES FOREARM (elbow flexed at 90°) *"Keeping your elbow bent and close to your side, turn your forearm over so that your palm faces up, then turn your forearm over so that your palm faces down"* **Note:** movement in one direction only = partial movement (score 1a or 1b).
				/2	12. CLOSES HAND FROM FULLY OPENED POSITION *"Make a fist, keeping your thumb on the outside"* **Note:** must extend wrist slightly (ie wrist cocked) to obtain full marks.
				/2	13. OPENS HAND FROM FULLY CLOSED POSITION *"Now open your hand all the way"*
				/2	14. OPPOSES THUMB TO INDEX FINGER (tip to tip) *"Make a circle with your thumb and index finger"*
				/2	15. FLEXES HIP IN SITTING *"Lift your knee as high as you can"*
				/2	16. EXTENDS KNEE IN SITTING *"Straighten your knee by lifting your foot up"*
				/2	17. FLEXES KNEE IN SITTING *"Slide your foot back under you as far as you can"* **Note:** start with affected foot forward (heel in line with toes of other foot).
				/2	18. DORSIFLEXES ANKLE IN SITTING *"Keep your heel on the ground and lift your toes off the floor as far as you can"*
				/2	19. PLANTARFLEXES ANKLE IN SITTING *"Keep your toes on the ground and lift your heel off the floor as far as you can"*
				/2	20. EXTENDS KNEE AND DORSIFLEXES ANKLE IN SITTING *"Straighten your knee and bring your toes towards you"* **Note:** extension of knee without dorsiflexion of ankle = partial movement (score 1a or 1b)
				/3	21. RISES TO STANDING FROM SITTING *"Stand up, try to take equal weight on both legs"* **Note:** pushing up with hand(s) to stand = aid (score 2); asymmetry such as trunk lean, trendelenburg, hip retraction, or excessive flexion or extension of the affected knee = marked deviation (score **1a** or **1c**).

SCORE 4	3	2	1		
				/3	**STANDING** 22. MAINTAINS STANDING FOR 20 COUNTS *"Stand on the spot while I count to twenty"*
				/2	**STANDING (holding onto a stable support to assist balance for items 23-25)** 23. ABDUCTS AFFECTED HIP WITH KNEE EXTENDED *"Keep your knee straight and your hips level, and raise your leg to the side"*
				/2	24. FLEXES AFFECTED KNEE WITH HIP EXTENDED *"Keep your hip straight, bend your knee back and bring your heel towards your bottom"*
				/2	25. DORSIFLEXES AFFECTED ANKLE WITH KNEE EXTENDED *"Keep your heel on the ground and lift your toes off the floor as far as you can"*
				/3	**STANDING AND WALKING ACTIVITIES** 26. PLACES AFFECTED FOOT ONTO FIRST STEP (or stool 18 cm high) *"Lift your foot and place it onto the first step (or stool) in front of you"* Note: returning the foot to the ground is not scored, use of handrail - aid (score 2)
				/3	27. TAKES 3 STEPS BACKWARDS (one and a half gait cycles) *"Take three average sized steps backwards, placing one foot behind the other"*
				/3	28. TAKES 3 STEPS SIDEWAYS TO AFFECTED SIDE *"Take three average sized steps sideways towards your weak side"*
				/3	29. WALKS 10 METERS INDOORS (on smooth, obstacle free surface) *"Walk in a straight line over to (a specified point 10 meters away)"* Note: orthotic = aid (score 2), longer than 20 seconds = marked deviation (score 1c)
				/3	30. WALKS DOWN 3 STAIRS ALTERNATING FEET *"Walk down three stairs, place only one foot at a time on each step if you can"* Note: handrail = aid (score 2), non-alternating feet = marked deviation (score 1a or 1c).

APPENDIX 9

Results from the Direct Observation Reliability Study

APPENDIX 9.1

Distribution of Scores Within Each Item

(N = 2 Raters x 20 Subjects = 40 Ratings)

ITEM	SCORE					
	0	1a	1b	1c	2	X
Upper Extremity:						
1.	10	3	5	1	19	2
2.	8	4		5	23	
11.	12	3	4	2	19	
12.	11	7		8	14	
13.	11	2		4	21	2
14.	11	6	1	2	20	
15.	12	7	3	2	16	
16.	10	8	4	18		
17.	12	3	2	3	18	2
18.	11	3	3	6	17	
19.	15	4	1	1	19	
20.	14	2		7	17	
21.	12	5	7	1	11	4
Lower Extremity:						
3.		1		4	35	
4.	6	4		8	22	
6.	1	12	3	8	14	2
22.	4	10	1	6	19	
23.	3	10	3	5	19	
24.	6	9	4	1	20	
25.	4	4	4		23	4
26.	12	8	1	2	17	
27.	3	14	3	2	18	
31.	5	11	1	3	18	2
32.	12	5	3	3	15	2
33.	11		1	2	21	5

0	1a	1b	1c	2	3	X
Basic Mobility:						
5.	2		13	1	24	
7.		3	4	2	31	
8.	2	5	2	2	29	
9.	4	1	11	3	21	
10.			1		39	
28.					40	
29.	8	1	11	6	14	
30.			12	6	22	
34. 4	5	1	12	2	16	
35. 2	7	1	8	5	15	2
36. 2	10		8	3	15	2
37. 2	7	1	6	4	16	4
38. 4	7	1	4	4	16	4
39. 2	12		3	5	16	2
40. 2	12		4	4	16	2
41. 4	8		12	4	10	2
42. 6	11		4	9	8	2
43. 6	11		4	9	8	2

APPENDIX 9.2

Inter-Item, Item to Total and Item to Subscale Correlations of the 43 Items in the Test Version of STREAM

ITEM:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	Item-total	Item-subscale
1	1.0																																											.88	.83
2	.75	1.0																																										.76	.83
3	.58	.42	1.0																																									.54	.50
4	.60	.52	.33	1.0																																									.73	.84
5	.60	.48	.33	.42	1.0																																								.63	.60
6	.58	.64	.57	.63	.53	1.0																																							.73	.71
7	.60	.35	.31	.60	.55	.53	1.0																																						.59	.60
8	.47	.56	.39	.62	.46	.45	.50	1.0																																					.60	.63
9	.72	.53	.44	.52	.67	.48	.51	.64	1.0																																					.82	.88
10	.46	.41	.46	.37	.07	.31	.46	.63	.49	1.0																																				.51	.57
11	.85	.83	.35	.74	.64	.69	.50	.60	.68	.40	1.0																																			.89	.86
12	.78	.83	.40	.69	.47	.54	.29	.34	.54	.26	.83	1.0																																		.81	.95
13	.68	.81	.51	.42	.45	.52	.30	.40	.45	.23	.70	.89	1.0																																	.73	.90
14	.80	.78	.43	.68	.55	.61	.35	.35	.68	.28	.85	.88	.75	1.0																																.87	.92
15	.83	.76	.42	.65	.52	.64	.33	.33	.56	.27	.83	.84	.73	.93	1.0																															.81	.89
16	.87	.79	.46	.62	.50	.52	.29	.39	.65	.31	.83	.91	.87	.93	.90	1.0																														.86	.95
17	.74	.81	.46	.51	.56	.57	.34	.34	.53	.19	.80	.91	.92	.88	.88	.90	1.0																													.78	.96
18	.65	.78	.48	.47	.48	.59	.36	.36	.46	.20	.72	.90	.95	.82	.79	.84	.95	1.0																												.74	.91
19	.79	.72	.48	.65	.59	.56	.27	.27	.57	.15	.79	.91	.89	.88	.85	.91	.96	.91	1.0																											.78	.95
20	.74	.72	.48	.39	.50	.48	.17	.27	.49	.15	.74	.91	.89	.79	.76	.86	.90	.85	.95	1.0																										.72	.91
21	.81	.69	.49	.58	.60	.58	.37	.43	.66	.24	.85	.88	.81	.84	.80	.87	.87	.82	.92	.92	1.0																									.85	.92
22	.80	.67	.52	.82	.55	.66	.58	.58	.74	.57	.81	.67	.52	.69	.66	.69	.58	.47	.57	.57	.70	1.0																								.85	.87
23	.82	.67	.59	.82	.43	.65	.55	.55	.67	.43	.79	.73	.57	.72	.70	.72	.61	.56	.60	.60	.71	.93	1.0																							.86	.89
24	.82	.72	.44	.75	.61	.69	.76	.65	.70	.55	.88	.70	.66	.72	.69	.73	.69	.65	.66	.61	.75	.84	.85	1.0																						.92	.92
25	.76	.64	.52	.74	.65	.64	.58	.49	.58	.44	.79	.77	.80	.82	.80	.83	.76	.80	.79	.74	.84	.70	.80	.86	1.0																					.90	.90
26	.69	.53	.35	.86	.55	.61	.69	.50	.64	.40	.81	.66	.54	.71	.68	.68	.68	.56	.60	.55	.75	.81	.79	.88	.91	1.0																				.86	.94
27	.80	.67	.52	.82	.50	.66	.64	.70	.74	.64	.81	.67	.55	.69	.66	.69	.58	.54	.57	.57	.72	.93	.93	.91	.83	.87	1.0																			.89	.94
28	.34	.41	.46	.17	.01	.21	.29	.40	.39	.68	.28	.38	.46	.40	.39	.42	.42	.44	.37	.37	.37	.36	.41	.43	.53	.40	.50	1.0																		.49	.47
29	.62	.67	.49	.63	.51	.71	.64	.64	.71	.47	.67	.66	.57	.71	.57	.61	.62	.65	.63	.54	.58	.59	.61	.65	.57	.53	.59	.41	1.0																	.79	.83
30	.79	.66	.62	.63	.66	.52	.31	.47	.82	.55	.63	.64	.58	.75	.64	.77	.62	.55	.65	.61	.66	.69	.72	.65	.64	.51	.69	.50	.77	1.0																.82	.84
31	.73	.51	.39	.79	.58	.71	.69	.64	.80	.51	.78	.52	.39	.67	.64	.63	.66	.62	.50	.39	.63	.81	.76	.85	.85	.89	.88	.39	.64	.63	1.0															.87	.90
32	.64	.67	.32	.83	.50	.68	.68	.47	.58	.38	.76	.58	.48	.65	.67	.62	.53	.49	.53	.47	.67	.77	.75	.85	.89	.97	.84	.38	.52	.69	.87	1.0														.82	.92
33	.76	.52	.44	.76	.60	.58	.60	.48	.61	.33	.76	.74	.72	.76	.74	.73	.71	.75	.77	.72	.87	.71	.83	.84	.91	.93	.83	.40	.54	.56	.83	.90	1.0													.87	.89
34	.75	.53	.42	.67	.59	.65	.60	.56	.88	.47	.74	.58	.48	.74	.62	.70	.55	.50	.59	.47	.64	.75	.73	.81	.74	.77	.79	.38	.79	.81	.91	.79	.73	1.0												.90	.96
35	.81	.62	.47	.67	.64	.70	.56	.61	.87	.53	.82	.66	.54	.73	.65	.72	.55	.53	.61	.53	.69	.82	.78	.85	.79	.82	.82	.39	.83	.80	.89	.78	.78	.93	1.0											.92	.97
36	.75	.64	.65	.60	.54	.69	.69	.57	.79	.58	.75	.67	.59	.71	.63	.70	.54	.57	.59	.51	.64	.74	.71	.80	.78	.79	.79	.58	.83	.77	.84	.75	.74	.88	.96	1.0										.89	.93
37	.82	.59	.69	.69	.65	.67	.60	.64	.88	.55	.84	.65	.53	.75	.67	.74	.57	.53	.64	.56	.73	.85	.80	.87	.86	.85	.84	.60	.78	.81	.91	.82	.87	.95	.98	.91	1.0									.95	.99
38	.81	.57	.62	.70	.66	.64	.62	.62	.89	.53	.81	.63	.49	.77	.70	.73	.57	.53	.62	.51	.70	.82	.77	.86	.84	.86	.84	.64	.77	.79	.94	.83	.87	.97	.96	.91	.97	1.0								.94	.99
39	.88	.66	.66	.66	.59	.64	.51	.59	.86	.56	.85	.73	.59	.81	.74	.80	.64	.59	.70	.62	.76	.81	.82	.88	.82	.81	.86	.67	.90	.85	.88	.78	.84	.95	.95	.92	.97	.97	1.0							.95	.97
40	.83	.63	.64	.67	.64	.65	.57	.65	.88	.54	.84	.66	.55	.76	.69	.76	.59	.54	.65	.57	.74	.81	.78	.88	.82	.84	.86	.66	.78	.82	.92	.81	.80	.96	.97	.93	.99	.99	.99	1.0						.94	.98
41	.66	.56	.66	.63	.47	.65	.53	.63	.73	.41	.64	.65	.53	.64	.53	.64	.47	.49	.51	.51	.61	.77	.76	.74	.71	.76	.77	.57	.75	.76	.73	.76	.72	.82	.87	.87	.83	.79	.81	.81	1.0					.82	.85
42	.73	.60	.63	.67	.55	.71	.60	.68	.82	.44	.74	.67	.52	.78	.69	.70	.58	.56	.61	.53	.68	.75	.76	.82	.79	.82	.81	.64	.82	.81	.84	.82	.78	.94	.90	.89	.90	.92	.90	.91	.91	1.0				.91	.94
43	.73	.60	.63	.67	.55	.71	.60	.68	.82	.44	.74	.67	.52	.78	.69	.70	.58	.56	.61	.53	.68	.75	.76	.82	.79	.82	.81	.64	.82	.81	.84	.82	.78	.94	.90	.89	.90	.92	.90	.91	.91	1.0	1.0			.91	.94

APPENDIX 9.3

The Total STREAM and Subscale Scores [a] Assigned to Twenty Subjects by Assessors and Observers

	TOTAL SCORE		U/E Subscale		L/E Subscale		Mobility Subscale	
	Assessor	Observer	Assessor	Observer	Assessor	Observer	Assessor	Observer
*	23	25	0	0	25	28	37	40
*	27	31	0	0	45	45	33	43
	29	32	65	75	0	0	23	23
*	29	36	0	0	50	70	33	37
	33	36	15	20	30	20	47	57
*	35	32	6	0	55	50	39	39
	39	39	60	55	25	30	33	33
	44	39	10	15	25	30	80	60
	54	59	40	40	50	50	33	43
*	56	56	95	95	33	33	43	43
*	72	75	72	83	100	100	53	53
	83	89	60	65	95	100	90	97
	83	83	90	80	65	75	90	90
	84	96	80	95	100	100	77	93
	89	89	60	65	100	100	100	97
	91	93	100	100	100	100	72	78
*	94	91	95	95	95	95	93	87
	100	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100

Note: scores are for 30 item STREAM

a Subscale totals are transformed to scores out of 100.

* Scores corrected for missing items (items scored as X).

APPENDIX 9.4

Indices of Agreement for the Qualitative Divisions Within Category '1' for Pairs of Raters Scoring Twenty Subjects

Item Number	Agreement (%) Crude	Expected	Kappa Statistic [a]	95% Confidence Interval for Kappa	N[b]
Upper Extremity Subscale:					
1	67	33	.50	-0.11--1.11	3
2	100	50	1.00	1.00--1.00	4
11	100	33	1.00	1.00--1.00	3
12	67	44	.40	-0.14--0.94	6
13	0	0	0.00[c]	0.00--0.00	1
15	75	50	.50	-0.23--1.23	4
17	50	31	.27	-0.31--0.86	4
18	80	36	.69	0.16--1.22	5
19	50	50	0.00[c]	0.00--0.00	2
20	100	63	1.00	1.00--1.00	4
Lower Extremity Subscale:					
4	60	52	.17	-0.71--1.04	5
22	50	39	.18	-0.15--0.51	8
23	75	41	.58	0.12--1.04	8
24	71	49	.44	-0.06--0.94	7
25	75	50	.50	-0.23--1.23	4
26	100	100	---[d]	---	4
27	89	56	.75	0.30--1.20	9
31	71	53	.39	-0.17--0.95	7
32	67	33	.50	-0.11--1.11	3
33	100	100	---[d]	---	1
Basic Mobility Subscale:					
5	100	76	1.00	1.00--1.00	7
8	50	38	.20	-0.61--1.01	4
9	88	53	.73	0.29--1.18	8
29	100	51	1.00	1.00--1.00	7
30	100	100	---[d]	---	5
34	88	48	.76	0.39--1.12	7
36	75	53	.47	-0.17--1.10	7
38	80	52	.58	-0.05--1.22	5
41	100	51	1.00	1.00--1.00	9
43	100	59	1.00	1.00--1.00	7

a Unweighted Kappa statistic

b N= the number of subjects (out of 20) scored '1' by both raters

c Numerator of 0, for item 13 there is only one rating in category '1'

d Kappa was undefined because the denominator was 0

APPENDIX 10

Results from the Videotaped Assessments Reliability Study

APPENDIX 10.1

Distribution of Scores Within Each Item for the Four Videotapes (N = 20 Raters x 4 Tapes = 80 Ratings)

ITEM	TIME 1 SCORE						TIME 2 SCORE					
	0	1a	1b	1c	2	(3)	0	1a	1b	1c	2	(3)
Upper Extremity:												
1.	22	3	28	14	13		17	14	35	1	13	
2.		3	3	51	23			2	1	44	33	
11.	28	16	15	2	19		24	19	20		17	
12.	26	38	2	13	1		25	37	3	13	2	
13.	28	40	10		2		26	44	9		1	
14.	28	32		2	18		26	34		2	18	
15.	26	38	12	1	3		25	38	11	2	4	
16.	25	49	6				20	57	2	1		
17.	5	64	6	4	1		4	66	7	2	1	
18.	24	30	4	18	4		16	35	6	16	7	
19.	53	7	3	8	9		46	15	4	8	7	
20.	25	38	1	13	3		27	36		15	2	
21.	59	18	3				59	20	1			
Lower Extremity:												
3.				2	78					1	79	
4.	16	22		3	39		16	16	3	2	43	
6.		27	4	27	22			26	1	26	27	
22.		20	2	30	28			18	5	23	34	
23.		44	11	4	21		1	30	24	2	23	
24.	6	28	11	11	24		4	31	12	10	23	
25.	25	26	9	2	18		23	20	15	1	21	
26.	24	27	6	2	21		26	21	5	3	25	
27.	3	47	10		20		3	44	13		20	
31.	14	27	2	13	24		9	33	4	7	27	
32.	29	28	4		19		25	26	7	2	20	
33.	44	13	4	1	18		41	13	7	1	18	
Basic Mobility:												
5.		2		67	3	8		1		59	11	9
7.		3	5	11		61		1	5	1	2	71
8.		3	2	7		68			8	1	13	58
9.		28	2	25	7	18		25	2	25	12	16
10.						80						80
28.				1		79					1	79
29.				35	25	20				22	37	21
30.		1		27	2	50		1	2	15	5	57
34.		28	3	26	3	20		24	4	22	10	20
35.				52	8	20		1		40	18	21
36.		23	1	36		20		20	2	31	7	20
37.		20		38	2	20		20		29	10	21
38.		20	3	35	2	20		18	1	33	8	20
39.		19	1	32	7	21		18	1	28	13	20
40.		20		33	6	21		18	1	28	13	20
41.				60		20				38	22	20
42.		34		26		20		30	3	24	3	20
43.		35		25		20		29	5	23	3	20

APPENDIX 10.1

Distribution of Scores Within Each Item for Videotape A

(N = 20 Raters)

ITEM	TIME 1 SCORE						TIME 2 SCORE					
	0	1a	1b	1c	2	(3)	0	1a	1b	1c	2	(3)
Upper Extremity:												
1.	2		15		3		2		10		8	
2.		1	3	4	12			1	1	5	13	
11.				1	19				4		16	
12.		4	2	13	1			3	3	13	1	
13.		8	10		2			10	9		1	
14.				2	18					2	18	
15.		4	12	1	3			3	11	2	4	
16.		14	6					18	1	1		
17.		9	6	4	1			11	6	2	1	
18.		1	3	12	4			1	2	11	6	
19.			3	8	9			1	4	8	7	
20.		4		13	3			3		15	2	
21.		17	3					19	1			
Lower Extremity:												
3.					20						20	
4.					20						20	
6.				1	19					1	19	
22					20						20	
23.					20						20	
24.					20						20	
25.			4	1	15				2		18	
26.					20						20	
27.					20						20	
31.				1	19						20	
32.			1		19						20	
33.			1	1	18			:		1	18	
Basic Mobility:												
5.				17	3					12	8	
7.						20						20
8.						20						20
9.				8	1	11				8	1	11
10.						20						20
28.						20						20
29.						20						20
30.						20						20
34.						20						20
35.						20						20
36.						20						20
37.						20						20
38.						20						20
39.						20						20
40.						20						20
41.						20						20
42.						20						20
43.						20						20

APPENDIX 10.1
Distribution of Scores Within Each Item for Videotape B
(N = 20 Raters)

ITEM	TIME 1 SCORE						TIME 2 SCORE					
	0	1a	1b	1c	2	(3)	0	1a	1b	1c	2	(3)
Upper Extremity:												
1.		3	13	1	3			6	9	1	4	
2.				14	6					12	8	
11.	3	7	10				4	8	8			
12.	20						20					
13.	20						20					
14.	20						20					
15.	20						20					
16.	20						20					
17.		20					1	18	1			
18.	17	3					13	7				
19.	19	1					20					
20.	20						20					
21.	20						20					
Lower Extremity:												
3.					20						20	
4.	16	4					16	3	1			
6.		20						19	1			
22.		1	1	16	2			4	2	12	2	
23.		14	2	3	1			9	8	1	2	
24.		8	6	6				9	6	5		
25.	20						20					
26.	5	13	2				9	11				
27.	1	18	1				1	17	2			
31.		1	2	12	5		2	3	2	6	7	
32.	10	10					10	10				
33.	20						20					
Basic Mobility:												
5.				19		1		1		17	1	1
7.				4		16				1		19
8.			2			18						20
9.				10	3	37				12	5	3
10.						20						20
28.						20					1	19
29.				15	5					11	9	
30.		1		11	1	7		1	2	12	5	
34.		1		19						19	1	
35.				18	2					14	6	
36.				20						18	2	
37.		1		18	1					17	3	
38.		3		17						19	1	
39.				15	4	1				15	5	
40.				15	4	1				15	5	
41.				20						19	1	
42.		2		18						20		
43.		1		19					1	19		

APPENDIX 10.1
Distribution of Scores Within Each Item for Videotape C
(N = 20 Raters)

ITEM	TIME 1 SCORE						TIME 2 SCORE					
	0	1a	1b	1c	2	(3)	0	1a	1b	1c	2	(3)
Upper Extremity:												
1.	20						15	2	3			
2.		1		17	2					15	5	
11.	10	7	2	1			8	6	5		1	
12.		20					1	18			1	
13.	3	17					1	19				
14.	2	18					2	18				
15.	1	19					2	18				
16.	3	17						20				
17.	1	19						20				
18.		14		6				12	2	5	1	
19	18	2					16	4				
20.	5	15					5	15				
21.	19	1					20					
Lower Extremity:												
3.				1	19						20	
4.				1	19						20	
6.		1		18	1			1		14	5	
22.		1	1	14	4					9	11	
23.		11	9					6	13		1	
24.		6	5	5	4			7	5	5	3	
25.		11	5	1	3			5	11	1	3	
26.		13	4	2	1			7	5	3	5	
27.		13	7					11	9			
31.	5	15					1	17	1	1		
32.		17	3				1	10	7	2		
33.	4	13	3				1	12	7			
Basic Mobility:												
5.				13		7				10	2	8
7.		2	3	1		14			2		1	17
8.						20				1	1	18
9.		11		7	2			8	1	4	5	2
10.						20						20
28.				1		19						20
29.				9	11					6	13	1
30.				10		10				2		18
34.		9	2	6	3			6	4	2	8	
35.				17	3					14	6	
36.		3	1	16				2	2	12	4	
37.				19	1			1		12	7	
38.			1	17	2				1	14	5	
39.				17	3					13	7	
40.				18	2					13	7	
41.				20						19	1	
42.		12		8				10	3	4	3	
43.		14		6				9	4	4	3	

APPENDIX 10.1
Distribution of Scores Within Each Item for Videotape D
(N = 20 Raters)

ITEM	TIME 1 SCORE						TIME 2 SCORE					
	0	1a	1b	1c	2	(3)	0	1a	1b	1c	2	(3)
Upper Extremity:												
1.				13	7			6	13		1	
2.		1		16	3			1		12	7	
11.	15	2	3				12	5	3			
12.	6	14					4	16				
13.	5	15					5	15				
14.	6	14					4	16				
15.	5	15					3	17				
16.	2	18						19	1			
17.	4	16					3	17				
18.	7	12	1				3	15	2			
19.	16	4					10	10				
20.		19	1				2	18				
21.	20						19	1				
Lower Extremity:												
3				1	19					1	19	
4.		18		2				13	2	2	3	
6.		6	4	8	2			6		11	3	
22.		18		2				14	3	2	1	
23.		19		1			1	15	3	1		
24.	6	14					4	15	1			
25.	5	15					3	15	2			
26.	19	1					17	3				
27.	2	16	2				2	16	2			
31.	9	11					6	13	1			
32.	19	1					14	6				
33.	20						20					
Basic Mobility:												
5.		2		18						20		
7.		1	2	6		11		1	3		1	15
8.		3		7		10			8		12	
9.		17	2		1			17	1	1	1	
10.						20						20
28.						20						20
29.				11	9					5	15	
30.				6	1	13				1		19
34.		18	1	1				18		1	1	
35.				17	3			1		12	6	1
36.		20						18		1	1	
37.		19		1				19				1
38.		17	2	1				18			2	
39.		19	1					18	1		1	
40.		20						18	1		1	
41.				20						20		
42.		20						20				
43.		20						20				

APPENDIX 10.1

Distribution of Agreement for Twenty Raters Scoring Tape A on Two Occasions

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1	1,2	0,2	
Upper Extremity:							
1	2	10	3	0	5	0	
2	0	4	9	0	7	0	
11	0	0	15	0	5	0	
12	0	19	1	0	0	0	
13	0	18	1	0	1	0	
15	0	15	2	0	3	0	
17	0	19	1	0	0	0	
18	0	14	4	0	2	0	
19	0	8	4	0	8	0	
20	0	17	2	0	1	0	
TOTAL:	2	124	42	0	32	0	200
		84.0 %			16.0 %		
Lower Extremity:							
4	0	0	20	0	0	0	
22	0	0	20	0	0	0	
23	0	0	20	0	0	0	
24	0	0	20	0	0	0	
25	0	2	15	0	3	0	
26	0	0	20	0	0	0	
27	0	0	20	0	0	0	
31	0	0	19	0	1	0	
32	0	0	19	0	1	0	
33	0	1	17	0	2	0	
TOTAL:	0	3	190	0	7	0	200
		96.5 %			3.5 %		

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	
Basic Mobility:										
5	0	11	0	2	0	0	0	0	7	
8	0	0	0	20	0	0	0	0	0	
9	0	6	1	9	0	0	0	0	4	
29	0	0	0	20	0	0	0	0	0	
30	0	0	0	20	0	0	0	0	0	
34	0	0	0	20	0	0	0	0	0	
36	0	1	0	19	0	0	0	0	0	
38	0	0	0	20	0	0	0	0	0	
41	0	0	0	20	0	0	0	0	0	
43	0	0	0	20	0	0	0	0	0	
TOTAL:	0	18	1	170	0	0	0	0	11	200
		94.5 %				0 %			5.5 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

APPENDIX 10.1 (cont)

Distribution of Agreement for Twenty Raters Scoring Tape B on Two Occasions

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1	1,2	0,2	
Upper Extremity:							
1	0	16	3	0	1	0	
2	0	12	6	0	2	0	
11	2	15	0	3	0	0	
12	20	0	0	0	0	0	
13	20	0	0	0	0	0	
15	20	0	0	0	0	0	
17	0	19	0	1	0	0	
18	11	1	0	8	0	0	
19	19	0	0	1	0	0	
20	20	0	0	0	0	0	
TOTAL	112	93	9	13	3	0	200
	92.0 %			8.0 %			
Lower Extremity:							
4	13	1	0	6	0	0	
22	0	17	1	0	2	0	
23	0	18	1	0	1	0	
24	0	20	0	0	0	0	
25	20	0	0	0	0	0	
26	3	9	0	8	0	0	
27	0	18	0	2	0	0	
31	0	10	3	1	5	1	
32	7	7	0	0	6	0	
33	20	0	0	0	0	0	
TOTAL	63	100	5	17	14	1	200
	84.0 %			15.5 %		0.5 %	

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	
Basic Mobility:										
5	0	17	0	0	0	1	0	0	2	
8	0	0	0	18	0	0	0	0	2	
9	0	10	2	2	0	0	4	0	2	
29	0	11	5	0	0	4	0	0	0	
30	0	9	0	3	0	1	0	0	1	
34	0	19	0	0	0	1	0	0	0	
36	0	1	0	19	0	0	0	0	0	
38	0	19	0	0	0	1	0	0	0	
41	0	19	0	0	0	1	0	0	0	
43	0	20	0	0	0	0	0	0	0	
TOTAL	0	125	7	42	0	9	4	0	13	200
	87.0 %				6.5 %			6.5 %		

Note: only the agreement on the 30 items that were retained (out of 43) is presented

APPENDIX 10.1 (cont)

Distribution of Agreement for Twenty Raters Scoring Tape C on Two Occasions

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1	1,2	0,2	
Upper Extremity:							
1	15	0	0	5	0	0	
2	0	4	1	0	5	0	
11	5	6	0	8	1	0	
12	0	18	0	1	1	0	
13	0	16	0	4	0	0	
15	1	18	0	1	0	0	
17	0	19	0	1	0	0	
18	0	19	0	0	1	0	
19	15	1	0	4	0	0	
20	2	12	0	6	0	0	
TOTAL	38	123	1	30	8	0	200
		81.0 %			19.0 %		
Lower Extremity:							
4	0	0	19	0	1	0	
22	0	7	2	0	11	0	
23	0	19	0	0	1	0	
24	0	14	1	0	5	0	
25	0	15	1	0	4	0	
26	0	15	1	0	4	0	
27	0	20	0	0	0	0	
31	1	15	0	4	0	0	
32	0	19	0	1	0	0	
33	0	15	0	5	0	0	
TOTAL	1	139	24	10	26	0	200
		82.0 %			18.0 %		

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	
Basic Mobility:										
5	0	8	0	5	0	2	0	0	5	
8	0	0	0	18	0	0	1	0	1	
9	0	13	2	0	0	3	0	0	2	
29	0	6	10	0	0	3	1	0	0	
30	0	2	0	10	0	0	0	0	8	
34	0	11	2	0	0	7	0	0	0	
36	0	16	0	0	0	4	0	0	0	
38	0	14	1	0	0	5	0	0	0	
41	0	19	0	0	0	1	0	0	0	
43	0	17	0	0	0	3	0	0	0	
TOTAL:	0	106	15	33	0	28	2	0	16	200
		77.0 %				15.0 %			8.0 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

APPENDIX 10.1 (cont)

Distribution of Agreement for Twenty Raters Scoring Tape D on Two Occasions

Item	Perfect Agreement			Disagreement by: One Category		Two Categories	
	0,0	1,1	2,2	0,1	1,2	0,2	
Upper Extremity:							
1	0	19	0	0	1	0	
2	0	12	2	0	6	0	
11	10	3	0	7	0	0	
12	4	14	0	2	0	0	
13	5	15	0	0	0	0	
15	2	14	0	4	0	0	
17	2	15	0	3	0	0	
18	3	13	0	4	0	0	
19	8	2	0	10	0	0	
20	0	18	0	2	0	0	
TOTAL	34	125	2	32	7	0	200
		80.5 %			19.5 %		
Lower Extremity:							
4	0	17	0	0	3	0	
22	0	19	0	0	1	0	
23	0	19	0	1	0	0	
24	2	12	0	6	0	0	
25	2	14	0	4	0	0	
26	16	0	0	4	0	0	
27	1	17	0	2	0	0	
31	5	10	0	5	0	0	
32	13	0	0	7	0	0	
33	20	0	0	0	0	0	
TOTAL	59	108	0	29	4	0	200
		83.5 %			16.5 %		

Item	Perfect Agreement				Disagreement by: One Category			Two Categories		
	0,0	1,1	2,2	3,3	0,1	1,2	2,3	0,2	1,3	
Basic Mobility:										
5	0	20	0	0	0	0	0	0	0	
8	0	7	0	9	0	0	0	0	4	
9	0	18	0	0	0	2	0	0	0	
29	0	5	9	0	0	6	0	0	0	
30	0	0	0	12	0	0	1	0	7	
34	0	19	0	0	0	1	0	0	0	
36	0	19	0	0	0	1	0	0	0	
38	0	18	0	0	0	2	0	0	0	
41	0	20	0	0	0	0	0	0	0	
43	0	20	0	0	0	0	0	0	0	
TOTAL	0	146	9	21	0	12	1	0	11	200
		88.0 %				6.5 %			5.5 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented

APPENDIX 10.1 (CONT)

Distribution of Agreement on the Qualitative Divisions Within Category '1' for Twenty Raters Scoring Tape A

Item	Perfect Agreement			Disagreement			
	1a	1b	1c	1a 1b	1b 1c	1a 1c	N*
Upper Extremity:							
1		10					10
2			2	1		1	4
11							0
12	3	2	13	1			19
13	7	8		3			18
15	3	10		1	1		15
17.	8	3	2	4	2		19
18	1	1	9		3		14
19		2	2	1	1		6
20	3		13			1	17
TOTAL	25	36	41	11	7	2	122
		83.6 %		9.0 %	5.7 %	1.6 %	
Lower Extremity:							
4							0
22							0
23							0
24							0
25		2					2
26							0
27							0
31							0
32							0
33				1			1
TOTAL:	0	2	0	1	0	0	3
		66.7 %		33.3 %			
Basic Mobility:							
5			11				11
8							0
9			6				6
29							0
30							0
34							0
36			1				1
38							0
41							0
43							0
TOTAL:	0	0	18	0	0	0	18
		100 %					

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

* The number of raters (out of 20) scoring '1' on both occasions.

APPENDIX 10.1 (cont)

Distribution of Agreement on the Qualitative Divisions Within Category '1' for Twenty Raters Scoring Tape B

Item	Perfect Agreement			Disagreement			N*
	1a	1b	1c	1a 1b	1b 1c	1a 1c	
Upper Extremity:							
1	3	9	1	3			16
2	4	6		4			14
11							0
12							0
13							0
15							0
17	18			1			19
18	1						1
19							0
20							0
TOTAL	26	15	1	8	0	0	50
	84.0 %			16.0 %			
Lower Extremity:							
4				1			1
22			10	1	2	4	17
23	7	1		7	1	2	18
24	4	3	3	5	1	4	20
25							0
26	7			2			9
27	15			3			18
31	1	1	6	1		1	10
32	7						7
33							0
TOTAL	41	5	19	20	4	11	100
	65.0 %			20.0 %	4.0 %	11.0 %	
Basic Mobility:							
5			16				16
8							0
9			10				10
29			11				11
30			8			1	9
34			18			1	19
36			18				18
38			16			3	19
41			19				19
43			18		1	1	20
TOTAL:	0	0	134	0	1	6	141
	95.0 %				0.7 %	4.3 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

* The number of raters (out of 20) scoring '1' on both occasions.

APPENDIX 10.1 (cont)

Distribution of Agreement on the Qualitative Divisions Within Category '1' for Twenty Raters Scoring Tape C

Item	Perfect Agreement			Disagreement			N*
	1a	1b	1c	1a 1b	1b 1c	1a 1c	
Upper Extremity:							
1							0
2			13			1	14
11	1	1		4			6
12	18						18
13	16						16
15	18						18
17	19						19
18	12		5	2			19
19							0
20							0
TOTAL	84	1	18	6	0	1	110
		93.6 %		5.5 %		0.9 %	
Lower Extremity:							
4							0
22			6		1		7
23	4	7		8			19
24	4	2	1	2	3	2	14
25	5	5		4	1		15
26	6	3	1	3		2	15
27	8	4		8			20
31	13			1		1	15
32	9	2		7		2	20
33	8	1		6			15
TOTAL	57	24	8	39	5	7	140
		63.6 %		27.9 %	3.6 %	5.0 %	
Basic Mobility:							
5			8				8
8							0
9	4		8	1			13
29			6				6
30			2				2
34	6	1	2	2			11
36	2	1	12	1			16
38		1	13				14
41			19				19
43	8		4	4		1	17
TOTAL:	20	3	74	8	0	1	106
		91.5 %		7.5 %		0 .9 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented.

* The number of raters (out of 20) scoring '1' on both occasions.

APPENDIX 10.1 (cont)

Distribution of Agreement on the Qualitative Divisions Within Category '1' for Twenty Raters Scoring Tape D

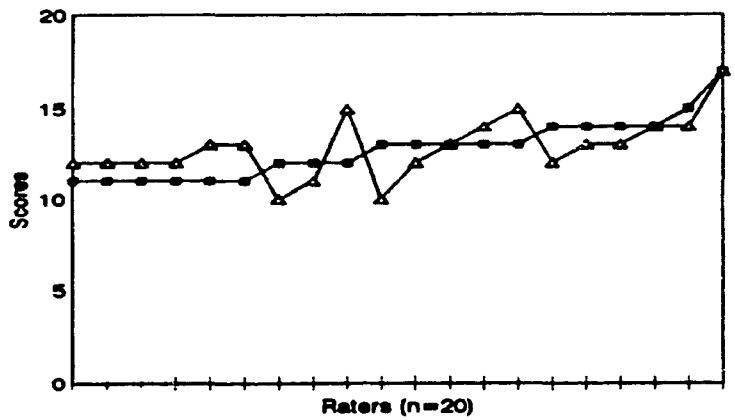
Item	Perfect Agreement			Disagreement			
	1a	1b	1c	1a 1b	1b 1c	1a 1c	N*
Upper Extremity:							
1	3	10		6			19
2	1		11				12
11	1			2			3
12	14						14
13	15						15
15	14						14
17	15						15
18	12	1					13
19	2						2
20	17			1			18
TOTAL	94	11	11	9	0	0	125
	92.8 %			7.2 %			
Lower Extremity:							
4	13		2	2			17
22	14		2	3			19
23	15		1	3			19
24	12						12
25	12			2			14
26							0
27	13			4			17
31	10						10
32							0
33							0
TOTAL	89		5	14	0	0	108
	87.0 %			13.0 %			
Basic Mobility:							
5			18			2	20
8		5		2			7
9	16	1		1			18
29			5				5
30							0
34	17		1	1			19
36	18					1	19
38	16			2			18
41			20				20
43	20						20
TOTAL	87	6	44	6	0	3	146
	93.8 %			4.1 %		2.1 %	

Note: only the agreement on the 30 items that were retained (out of 43) is presented

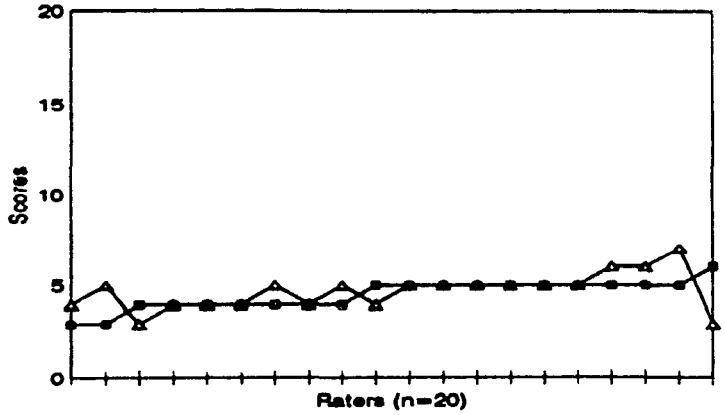
* The number of raters (out of 20) scoring '1' on both occasions.

APPENDIX 10.2
Videotaped Assessments Reliability Study:
Comparison of the Two Ratings on STREAM Upper Extremity Subscales

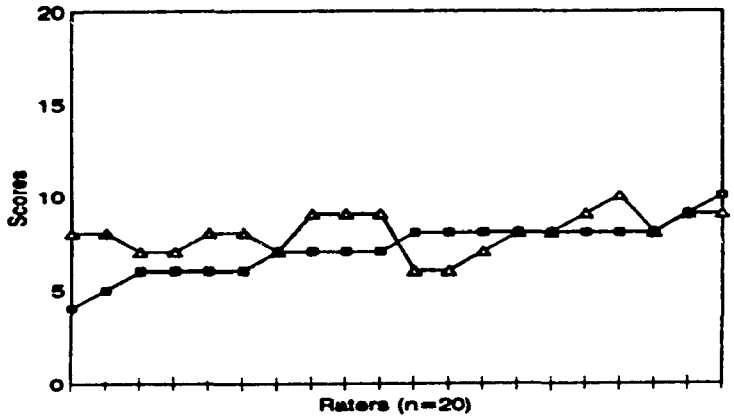
TAPE A



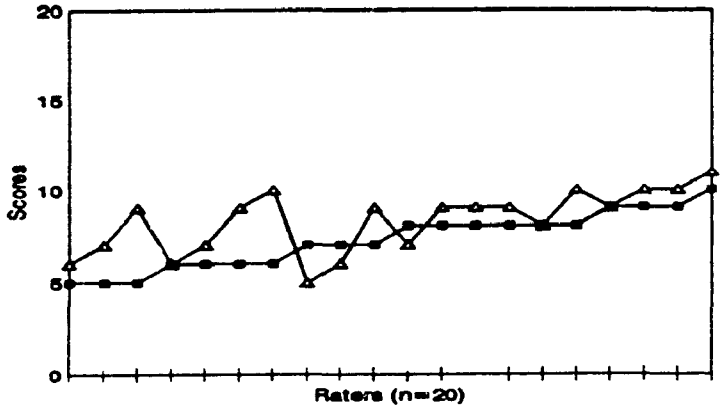
TAPE B



TAPE C



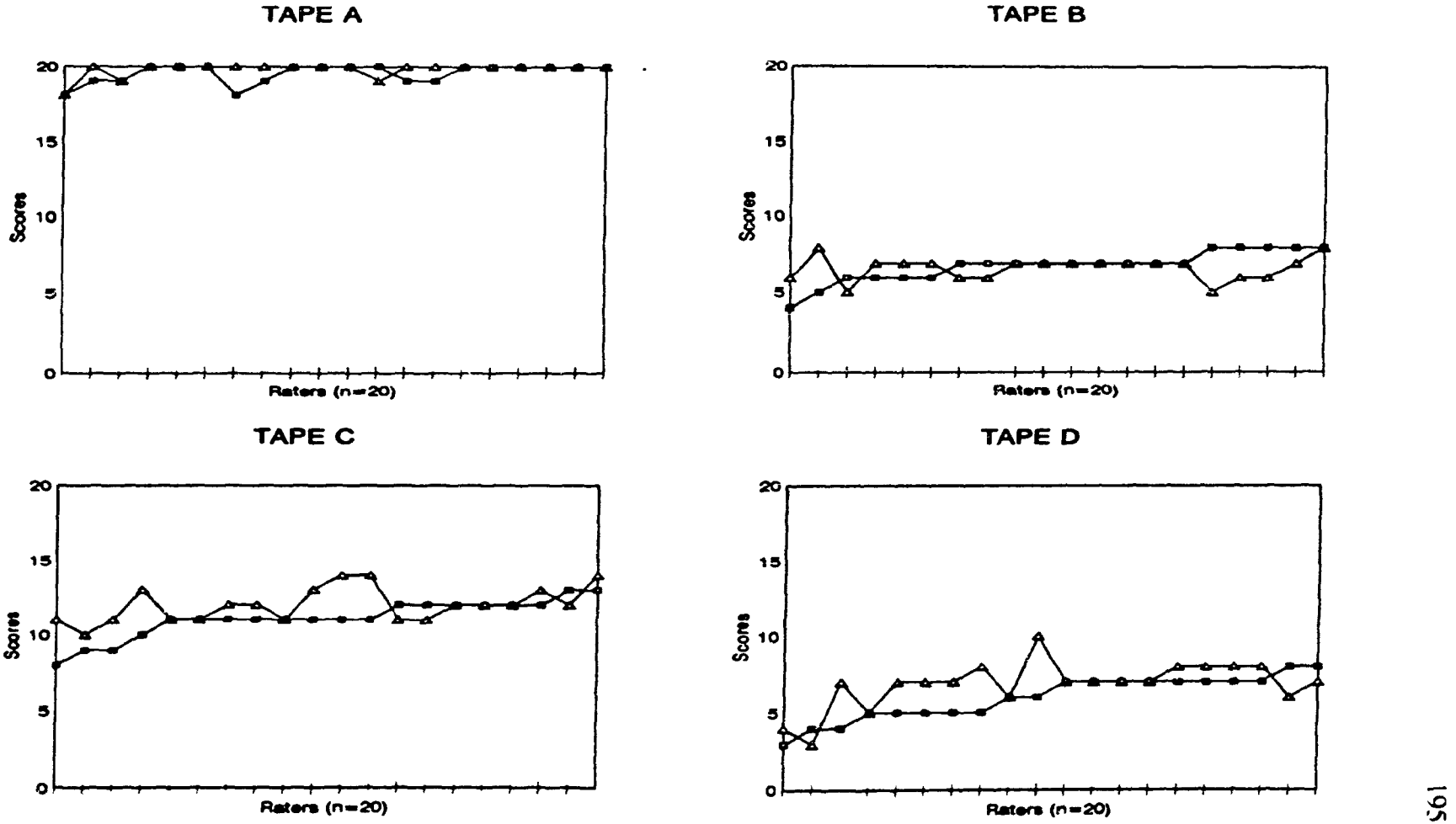
TAPE D



Key: ☐ Time 1 △ Time 2

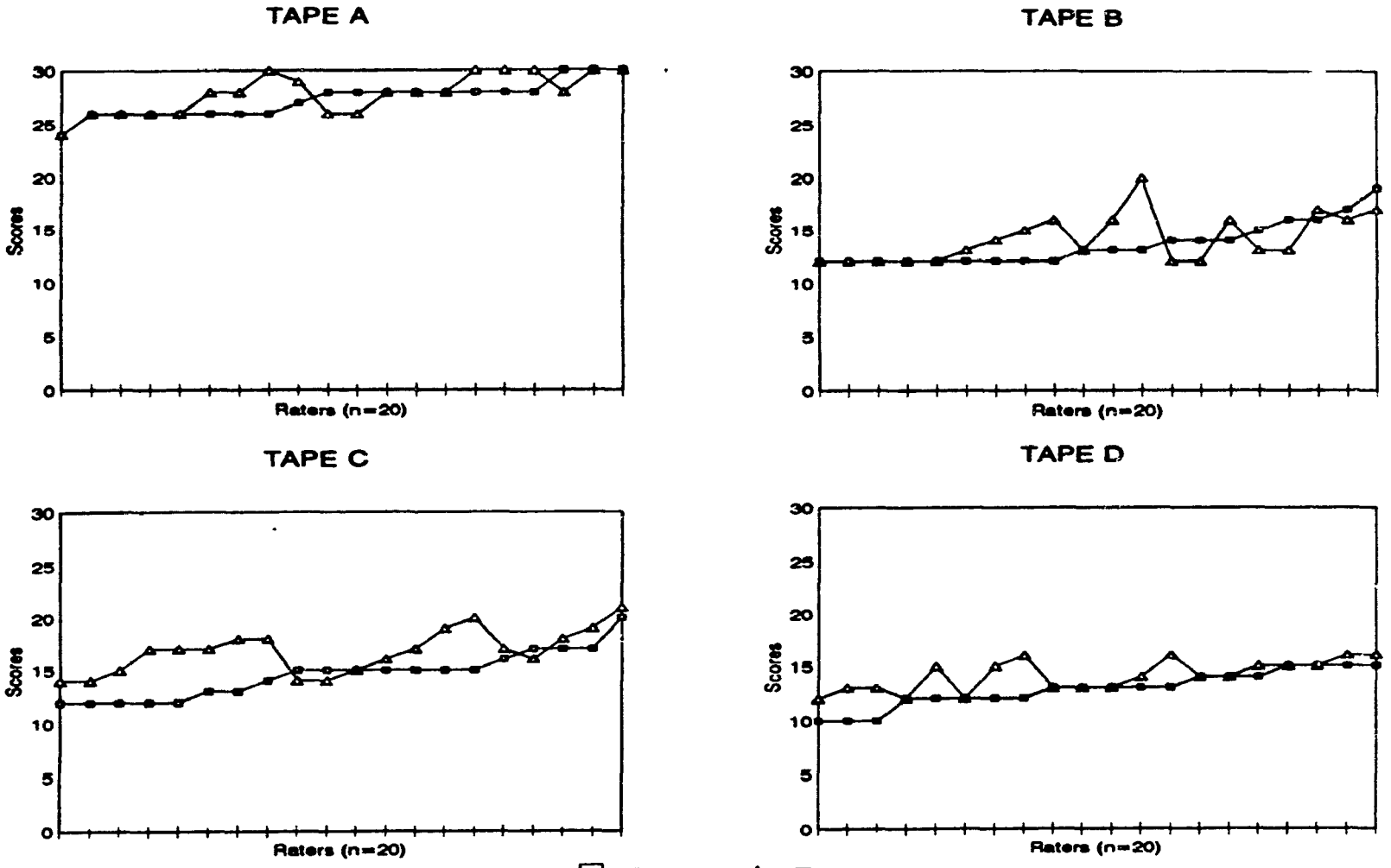
Note: GCC for upper extremity = .963 (model = subjects and occasions fixed, raters random)

APPENDIX 10.2
Videotaped Assessments Reliability Study:
Comparison of the Two Ratings on STREAM Lower Extremity Subscales



Note: ICC for lower extremity = .999 (model = subjects and occasions fixed, raters random)

APPENDIX 10.2
Videotaped Assessments Reliability Study:
Comparison of the Two Ratings on STREAM Basic Mobility Subscales



Note: GCC for basic mobility = .999 (model = subjects and occasions fixed, raters random)

APPENDIX 10.3

Assessment of Tape A by Twenty Raters on Two Occasions

Rater	Time One Score				Time Two Score			
	U/E	L/E	Mobility	TOTAL	U/E	L/E	Mobility	TOTAL
1.	11	18	26	55	12	18	26	56
2.	11	18	26	56	13	20	30	63
3.	11	19	26	56	12	19	26	57
4.	13	19	24	57	13	20	24	57
5.	11	20	26	57	12	20	28	60
6.	12	20	26	58	10	20	28	58
7.	12	20	28	58	11	20	30	61
8.	13	18	27	59	15	20	29	64
9.	11	19	28	59	12	20	30	62
10.	14	20	26	60	13	20	26	59
11.	12	20	28	60	15	20	28	63
12.	14	20	26	60	13	19	26	58
13.	13	20	28	60	14	20	30	64
14.	11	19	30	60	13	20	28	61
15.	13	19	28	61	10	20	28	58
16.	14	20	28	62	14	20	28	62
17.	14	20	28	62	12	20	26	58
18.	13	20	30	63	12	20	30	62
19.	17	20	28	65	17	20	26	63
20.	15	20	30	65	14	20	30	64

Note: scores are for 30 item STREAM; maximum total STREAM score is 70

APPENDIX 10.3 (cont)

Assessment of Tape B by Twenty Raters on Two Occasions

Rater	Time One Score				Time Two Score			
	U/E	L/E	Mobility	TOTAL	U/E	L/E	Mobility	TOTAL
1.	4	6	12	22	4	5	16	25
2.	3	7	12	22	4	7	15	26
3.	6	4	13	23	3	6	16	25
4.	4	7	12	23	4	7	12	23
5.	3	8	12	23	5	5	12	22
6.	5	5	14	24	5	8	16	29
7.	4	6	14	24	5	7	12	24
8.	5	7	12	24	6	6	12	24
9.	4	8	12	24	4	6	12	22
10.	4	7	13	24	3	6	20	29
11.	5	7	12	24	5	7	13	25
12.	5	8	12	25	5	7	12	24
13.	4	7	14	25	5	7	12	24
14.	5	7	13	25	5	7	13	25
15.	5	8	12	25	7	6	14	27
16.	5	7	15	27	5	7	13	25
17.	5	6	16	27	4	7	13	24
18.	4	8	16	28	4	8	17	29
19.	5	7	17	29	6	7	16	29
20.	5	6	19	30	5	7	17	29

Note: scores are for 30 item STREAM; maximum total STREAM score is 70.

APPENDIX 10.3 (cont)

Assessment of Tape C by Twenty Raters on Two Occasions

Rater	Time One Score				Time Two Score			
	U/E	L/E	Mobility	TOTAL	U/E	L/E	Mobility	TOTAL
1.	7	9	12	28	9	11	17	37
2.	6	11	12	29	8	11	14	33
3.	4	8	17	29	8	11	16	35
4.	7	9	15	31	9	10	19	38
5.	6	12	13	31	8	12	17	37
6.	8	11	12	31	8	13	14	35
7.	8	11	12	31	10	14	15	39
8.	8	11	12	31	6	12	17	35
9.	7	11	14	32	7	14	18	39
10.	8	11	13	32	9	12	18	39
11.	8	10	15	33	7	13	17	37
12.	6	12	16	34	7	13	17	37
13.	6	11	17	34	7	11	18	36
14.	8	12	15	35	8	11	14	33
15.	8	12	15	35	6	12	20	38
16.	7	13	15	35	9	14	14	37
17.	8	12	15	35	8	11	16	35
18.	5	13	17	35	8	12	19	39
19.	9	12	15	36	9	12	15	36
20.	10	11	20	41	9	11	21	41

Note: scores are for 30 item STREAM; maximum total STREAM score is 70

APPENDIX 10.3 (cont)

Assessment of Tape D by Twenty Raters on Two Occasions

Rater	Time One Score				Time Two Score			
	U/E	L/E	Mobility	TOTAL	U/E	L/E	Mobility	TOTAL
1.	5	3	13	21	9	4	14	27
2.	6	4	12	22	6	3	12	21
3.	5	5	13	23	7	7	13	27
4.	8	5	10	23	9	5	12	26
5.	7	7	10	24	9	8	13	30
6.	6	7	12	25	9	8	15	32
7.	8	4	13	25	7	7	13	27
8.	8	7	10	25	9	7	13	29
9.	5	7	13	25	6	7	13	26
10.	6	5	15	26	7	7	15	29
11.	6	8	12	26	10	7	16	33
12.	7	5	14	26	6	8	14	28
13.	8	5	14	27	9	7	15	31
14.	8	7	12	27	10	7	12	29
15.	7	7	13	27	5	7	16	28
16.	8	6	15	29	8	10	16	34
17.	10	7	12	29	11	8	15	34
18.	9	6	15	30	10	6	15	31
19.	9	7	15	31	9	8	16	33
20.	9	8	14	31	10	6	14	30

Note: scores are for 30 item STREAM; maximum total STREAM score is 70

APPENDIX 10.4
Summary of Signed Ranks for Items for Twenty Raters Scoring Four Videotapes on Two Occasions

Item Number	Signed Rank (P) TAPE A	TAPE B	TAPE C	TAPE D
Upper Extremity Subscale:				
1	-7.5 (053)	-0.5 (1.00)	-7.5 (.063)	-0 5 (1 00)
2	-2 (1 00)	-1.5 (.500)	-4.5 (.375)	-7 (.219)
11	4.5 (.375)	1 (1.00)	-7 5 (508)	-6 (453)
12	---	---	0 (1 00)	-1 5 (500)
13	0 5 (1.00)	---	-2.5 (.625)	---
15	-1 (1.00)	---	0.5 (1.00)	-2 5 (.625)
17	---	0.5 (1 00)	-0.5 (1.00)	-1 (1 00)
18	-1.5 (.500)	-9 (.289)	-0.5 (1.00)	-5 (.125)
19	4.5 (.727)	0.5 (1.00)	-2 5 (.625)	16 5 (.109)
20	0.5 (1.00)	---	0 (1 00)	1 5 (500)
Lower Extremity Subscale:				
4	---	---	-0.5 (1.00	-3 (250)
22	---	0 (1.00)	-21 (.065)	-0 5 (1 00)
23	---	-0.5 (1.00)	-0 5 (1 00)	0 5 (1.00)
24	---	---	1.5 (1 00)	-3 5 (688)
25	-3 (.250)	---	0 (1 00)	-2 5 (625)
26	---	9 (.289)	-5 (125)	-2 5 (625)
27	---	0 (1 00)	---	0 (1 00)
31	-0.5 (1 00)	0 (1.00)	-5 (125)	-4 5 (375)
32	-0 5 (1 00)	0 (1.00)	0 5 (1.00)	-10 (.125)
33	0.0 (1 00)	---	-4 5 (.375)	---
Basic Mobility Subscale:				
5	-10 (125)	-0.5 (1.00)	-4 (531)	---
8	---	-1 5 (.500)	1.5 (500)	-2.5 (625)
9	0 0 (1.00)	8 (.156)	-7.5 (063)	0 (1 00)
29	---	-5 (.125)	-5 (.125)	-10 5 (031)
30	---	-3 (.727)	-18 (008)	-13 (.070)
34	---	-0.5 (1.00)	-10 (.125)	-0.5 (1.00)
36	---	-1.5 (500)	-5 (125)	-0 5 (1 00)
38	---	-0.5 (1 00)	-4.5 (.375)	-1.5 (.500)
41	---	-0.5 (1.00)	-0 5 (1 00)	---
43	---	---	-3 (.250)	---