Measuring Physical Disability Following Limb Preservation for Lower Extremity Sarcoma

by

Aileen Mary Davis, BScPT MSc

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Medical Sciences
University of Toronto

© Copyright by Aileen M. Davis 1997
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.
ABSTRACT

Sarcoma is a life-threatening disease in which patients with extremity tumours undergo surgery with or without the use of adjuvant chemotherapy or irradiation to achieve local and systemic disease control. The goal of surgery is complete removal of the tumour to prevent local recurrence while preserving a functional limb. The consequence of treatment is that patients become physically disabled in order to potentially survive their disease. An outcome measure is required for monitoring patients' status over time and to evaluate different treatments within the context of clinical trials. The Toronto Extremity Salvage Score (TESS) is a measure of physical disability and was designed to serve these purposes. Previous work has shown that the TESS is feasible in a clinical setting and that it has test-retest reliability adequate for both group and individual patient use. The purpose of this thesis was to validate the TESS, particularly to confirm the content of the measure and to evaluate construct and criterion validity and the ability of the measure to detect change in status over time. Subjects completed the TESS and SF-36 pre-operatively and at 6 weeks, 3, 6 and 12 months post-operatively. The Musculoskeletal Tumor Society Rating Scale (MSTS) was completed by the clinician.

Open ended questions did not result in any items being added to the TESS nor were any items eliminated. For the TESS, Cronbach's alpha was high at 0.94 but factor analysis did not inform the item structure of the measure. A summary score was,
therefore, calculated based on clinimetric criteria. The TESS was shown to have construct validity by anticipated relationships with the SF-36 subscales and MSTS scores. Criterion validity was demonstrated by the ability of the TESS scores to differentiate between subjects with bone versus soft tissue tumours and amongst subjects based on the type of gait aid used. Over time, the TESS scores produced the anticipated pattern showing responsiveness to time and clinical events. The effect sizes detected were moderate. The current study has confirmed the content of the TESS and demonstrated its validity and responsiveness. The TESS is a useful measure of physical disability for evaluating treatment effectiveness in clinical trials, educating patients about their expected recovery and in monitoring individual patients.
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Jack Williams, who has been a true mentor. He provided continuous support, encouragement and intellectual stimulation; and he wondered why I always had another question!

To Bob Bell, words are inadequate to express the impact that your vision and enthusiasm have had on me over the past years. For all the opportunities and caring, thank you.

I would also like to thank the other members of my committee, Dr. Liza Badley and Dr. Karen Yoshida, for their thoughtful comments throughout my research and for their tireless review of my thesis.

This work would not have been possible without the assistance of Anthony Griffin who collected all the data and the encouragement of my colleagues, Zoe, Krista, Jay and David.

Nancy, Deirdre and Dorcas, your friendship and moral support helped make all of this possible. Finally, I thank my family for their patience and caring.

I would like to acknowledge the financial support of the National Cancer Institute of Canada and the National Cancer Institute of Canada-Clinical Trials Group for their grant awards in completing this work.
# TABLE OF CONTENTS

## Chapter One
Introduction  
1

## Chapter Two
### Foundations of the Toronto Extremity Salvage Score (TESS)
- Introduction  
- The TESS as a Measure of Disability  
- Overview of Measurement Theory  
  - Purpose of Measures  
  - Group Versus Individual Measurement  
  - Types of Measures  
  - Methods of Measure Development  
- Existing Measures of Health Status Used in Cancer Populations  
  - Generic Measures  
  - Cancer-specific Measures  
  - Disease-specific Measures  
- Development and Preliminary Evaluation of the Measurement Properties of the TESS  
37

## Chapter Three
### Methods of the Measurement Study
  - Study Setting  
  - Patient Sample  
  - Sample Size and Power Calculations  
  - Administration of the Instruments  
  - Instruments  
  - Analysis  
    - Data Quality  
    - Descriptive Statistics  
  - Developmental Study  
    - Item Generation and Reduction  
    - Internal Consistency and Factor Analysis  
      - Coding of the TESS "not applicable" Responses  
      - Internal Consistency  
      - Factor Analysis  
  - Validity Study  
    - Construct Validity  
    - Criterion Validity  
43

## Chapter Four
### Results
  - Introduction  
  - Patient Sample  
  - Power Calculations  
  - Developmental Study  
    - Item Generation and Reduction  
    - Internal Consistency and Factor Analysis  
      - Coding of the TESS "not applicable" Responses  
      - Internal Consistency  
      - Factor Analysis  
61
Chapter Four cont.

Validity Study 78
Construct Validity 78
Criterion Validity 80

Chapter Five
Discussion 94

Properties of the TESS 95
Comparing the TESS to Other Measures 101
The TESS as a Measure of Disability 104
Advantages and Limitations of the Thesis 108
Future Directions and Conclusions 110

Appendices 112

Bibliography 176
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Subjects Lost from the Study</td>
<td>62</td>
</tr>
<tr>
<td>Table 2</td>
<td>Lower Extremity Sample Characteristics</td>
<td>63</td>
</tr>
<tr>
<td>Table 3</td>
<td>Percent Difficulty Ratings and Totally Unimportant Ratings for Each TESS Item at Six Months Post-operative</td>
<td>67</td>
</tr>
<tr>
<td>Table 4</td>
<td>Correlation Matrix at Six Months Post-operatively</td>
<td>70</td>
</tr>
<tr>
<td>Table 5</td>
<td>Lower Extremity Six Months Post-operatively: Summary of the Correlation Matrix</td>
<td>72</td>
</tr>
<tr>
<td>Table 6</td>
<td>Principal Component Analysis: Orthogonal Rotation Six Months Post-operatively</td>
<td>76</td>
</tr>
<tr>
<td>Table 7</td>
<td>Lower Extremity Sub-groups: Principal Component Analysis</td>
<td>77</td>
</tr>
<tr>
<td>Table 8</td>
<td>Correlation Matrix of Measures</td>
<td>79</td>
</tr>
<tr>
<td>Table 9</td>
<td>Mean Post-operative TESS Scores by Use of Ambulatory Aid</td>
<td>84</td>
</tr>
<tr>
<td>Table 10</td>
<td>Relative Efficiency of the Toronto Extremity Salvage Score</td>
<td>93</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1  Magnetic resonance image of an osteosarcoma of the distal femur (figure 1a); the bony defect in the distal femur has been reconstructed with a metallic prosthesis (figure 1b) 5

Figure 2  Magnetic resonance image of a soft tissue sarcoma of the anteromedial thigh (figure 2a); intra-operative picture of the tumour resection (figure 2b) 6

Figure 3  Study Schema 44

Figure 4  Discrimination of TESS Scores and Disability 81

Figure 5  Boxplot of the Toronto Extremity Salvage Score Over Time 85

Figure 6  Boxplot of the SF-36 Over Time 86

Figure 7  Proximal Thigh Soft Tissue Tumour Subjects: TESS Scores Over Time 89

Figure 8  Distal Thigh Soft Tissue Tumour Subjects: TESS Scores Over Time 89

Figure 9  Subjects Reconstructed with Prostheses: TESS Scores Over Time 91

Figure 10  Subjects Reconstructed by Curettage and Bone Graft: TESS Scores Over Time 91
LIST OF APPENDICES

Appendix 1  Review of Disability Models  113
Appendix 2  Development of the SF-36  129
Appendix 3  Toronto Extremity Salvage Score Questionnaire  144
Appendix 4  Factor Analysis Results Over Time  171

ix
CHAPTER ONE

INTRODUCTION

Sarcoma is a life threatening disease with a sixty to seventy percent five year survival rate (Malawar, 1993; McLaughlin, 1995). However, those patients who do survive may experience significant disability as a result of the treatment for their disease (Lewis, 1992; Malawar, 1993). Consequently, the traditional indicators of outcome, disease activity and mortality data provide only a partial evaluation of the outcomes of treatment for this population. The short-fall of traditional outcome measures is common in modern medicine as large segments of the population seek treatment for chronic disease (Albrecht, 1992; Pope, 1991). Recognition of the limitations of mortality and morbidity data in evaluating chronic diseases has resulted in the use of health status or quality of life measures to evaluate treatment effectiveness and guide treatment decisions (Bergner, 1989; Feinstein, 1987; Ware, 1991). In particular, these measures provide assessment of treatment effectiveness based on the patient's perception of his or her health (Rineberg, 1990). In order to describe the disability experienced by the sarcoma population and to evaluate new treatment interventions, a relevant outcome measure with proven measurement properties is required.

Sarcomas are malignant tumours of mesenchymal origin (i.e. the connective tissues such as muscle, fat, bone) and occur most frequently in the extremities (Storm, 1994). The incidence of soft tissue sarcoma is twenty to thirty per million with approximately sixty percent occurring in the extremities (McLaughlin, 1995; Storm, 1994). Soft tissue sarcomas are most common in the fourth decade of life and the
incidence is slightly higher in males (Storm, 1994). The incidence of malignant, extremity bone tumours is two to three per million (Malawar, 1993). Bone tumours vary in age of onset by pathological sub-type with osteosarcoma and Ewing's sarcoma most common in the second and third decade and chondrosarcoma more common in the fifth and sixth decade. Again there is a slightly greater propensity for bone tumours to occur in males.

Patients with soft tissue tumours typically present with a recently noticed asymptomatic mass, or a sudden increased size in a long-standing asymptomatic mass. Patients with bone tumours more frequently present complaining of pain. The natural history of untreated extremity sarcoma involves progressive growth of the local tumour and eventual systemic spread to the lungs and, less commonly, the skeleton. Diagnosis and staging includes both tissue sampling by biopsy and testing for systemic disease using chest computed tomography and total body bone scan. The extent of local disease is determined by magnetic resonance imaging. Based on the results of these investigations treatment planning occurs. The consequences of no treatment in this disease are catastrophic both functionally and in survival.

The goal of treatment for patients with disease limited to the local site of origin is complete removal of the original tumour with preservation of the functional extremity. The function of the preserved limb is dependent on the amount of tissue that must be excised in order to completely remove the tumour and the surgical options that are available to reconstruct the defect. Consequently, the extremity sarcoma population may experience significant physical disability in order to potentially survive their disease (Bell, 1989; Bell, 1991; Capanna, 1991; Karasek, 1992; Roberts, 1991). Since the incidence of sarcoma is highest in those aged twenty to fifty years, the disease affects individuals who are both active and economically productive. Standardized measures of extremity function are necessary to evaluate the impact of treatment.

The basic classification and treatment of bone and soft tissue sarcomas is based on the tissue of origin, histological features and pathological grade of the tumour. Sarcomas
may demonstrate differentiated morphological characteristics of mesenchymal tissue (e.g. osteosarcoma produces bone, chondrosarcoma produces cartilage), or may not resemble any normal histological structure (e.g. malignant fibrous histiocytoma bears no resemblance to normal tissue). The grade of the tumour is determined by the usual features observed under the microscope (mitoses, nuclear aberrance and pleomorphism as well as macroscopic and microscopic necrosis). Cancer outcome in both bone and soft tissue sarcoma is related to tumour grade, size and whether metastases are evident at diagnosis (Bramwell, 1994; Lewis, 1992; Link, 1991; Pisters, 1996). Other determinants of outcome (e.g. percent necrosis of osteosarcoma in response to chemotherapy) may be specific to individual tumour types (Davis, 1994).

The natural history and treatment of most soft tissue sarcoma histological subtypes are similar. However, bone tumour sub-typing is extremely important. Osteosarcoma and Ewing's sarcoma patients are assumed to have micro-metastases at the time of diagnosis and, without chemotherapy, five year survival is less than twenty percent, whereas chemotherapy in conjunction with surgery results in survival of greater than fifty percent (Jaffe, 1974; Link, 1991; Rosen, 1974). Five year survival rates for soft tissue sarcoma are sixty-five percent (McLaughlin, 1995), for osteosarcoma sixty percent and for chondrosarcoma (which tends to have a significantly lower risk of metastases) ninety percent (Malawar, 1993).

The basic treatment protocol for sarcoma requires surgical resection of the primary tumour with or without local and/or systemic adjuvant therapy. The treatment protocols are determined by the tumour type. For example, the survival of patients with osteosarcoma or Ewing's sarcoma has been dramatically improved by the use of adjuvant chemotherapy (Jaffe, 1974; Link, 1991; Rosen, 1974). Other bone tumours (e.g. chondrosarcoma) are not proven to be responsive to chemotherapy.

Chemotherapy is not recognized as effective in improving systemic outcome in adult soft tissue sarcoma subtypes (Bramwell, 1994) and adjuvant treatment (irradiation)
in this disease usually involves treatment designed to permit limb preservation rather than amputation surgery. Since the advent of modern imaging and adjuvant therapies, amputation of the limb is rarely necessary and limb preservation is a viable option in over ninety percent of patients with bone or soft tissue sarcoma (Lewis, 1992; Simon, 1986).

In planning local treatment of both bone and soft tissue sarcoma, clinicians recognize the importance of the "pseudocapsule". The pseudocapsule surrounding the periphery of the extremity sarcoma is comprised of compressed normal muscle and fat that is infiltrated by both invading tumour cells and reactive inflammatory cells. Because of the presence of tumour cells in this pseudocapsule, it is critical that this layer be removed with the tumour to prevent local recurrence (Bell, 1989; Davis, in press). In order to ensure total removal, a layer of normal tissue that covers the pseudocapsule is resected with the tumour when possible.

Soft tissue sarcomas require complete surgical removal with or without the use of external beam irradiation. Radiotherapy damages the DNA of the tumour cells, decreases their ability to grow in the wound and potentially causes cell death. Surgery combined with radiotherapy allows preservation of critical structures such as nerves and blood vessels as well as the removal of a smaller volume of normal tissue.

After removal of a bone tumour, the skeletal defect is reconstructed. In selecting a reconstruction technique appropriate to the bony anatomical defect, the surgeon chooses from several options including autogenous bone graft, bone transplant or prosthetic reconstruction (Enneking, 1986; Lewis, 1992). Defects following soft tissue sarcoma removal usually are closed by suturing the layers of the remaining soft tissues together. However, tissue transfers and/or skin grafts may be required to close the wound. Figures 1 and 2 show examples of tumours and the surgical treatment.
Figure 1a: Magnetic resonance image of a soft tissue sarcoma filling the quadriceps compartment. The tumour is bright (white) compared to the normal muscle.

Figure 1b: Intraoperative view of the excision of a soft tissue sarcoma of the anterior thigh.
Figure 2a: Magnetic resonance image of an osteosarcoma of the distal femur. The normal marrow fat has a bright (white) signal whereas the marrow replaced by tumour is black.

Figure 2b: This x-ray shows the endoprosthesis used to reconstruct the distal femur following resection of an osteosarcoma.
The primary disabilities caused by treatment of tumours of the soft and bony tissues of the extremities relate primarily to the domain of physical function (Bell, 1989; Bell, 1991; Capanna, 1991; Karasek, 1992; Roberts, 1991). Therefore, in order to evaluate the consequences of extremity sarcoma and its treatment, the physical disability experienced by this population requires quantification. Physical function may be evaluated by clinical measures (e.g. range of motion of a joint), physiological-based measures (e.g. timed fifty foot walk test) and also by health status or quality of life questionnaires. Health-related quality of life is a multi-dimensional concept that includes the dimensions of physical, emotional and social function (Liang, 1987; Ware, 1991). These instruments may be generic or disease-specific. Physical function includes activity limitations such as restrictions in body movement (e.g. bending over), restrictions in mobility (e.g. walking, driving a car), restrictions in self-care (e.g. bathing) and restrictions in performing daily tasks and routines (Patrick, 1993). Any measure evaluating this construct should include these concepts.

A further requirement of a functional outcome measure for the extremity sarcoma population is that it be useful for the many heterogeneous subgroups within the population. Significant heterogeneity in a sarcoma population results from patient demographics, tumour type and the surgical options for reconstruction. Although there are ages of peak incidence, the age of patients treated in adult musculoskeletal oncology ranges from the teens to otherwise healthy octogenarians. As previously noted, surgical management may be combined with irradiation or chemotherapy, each of which may produce morbidity, (e.g. soft tissue fibrosis, cardiac myopathy, peripheral neuropathy). Consequently, the experience of treatment may vary depending on the modalities required to achieve the optimal chance for local and systemic disease control. The surgery necessary for tumour removal contributes to the heterogeneity of outcomes because of the variability in the extent of normal tissues excised and the variety of available reconstructive
techniques. In evaluating outcomes for the extremity sarcoma population, recognition of the heterogeneity of the population is an important consideration.

The existing clinical and physiological-based measures, including the Musculoskeletal Tumor Society Rating Scale (Enneking, 1987; Enneking, 1993), neither provide an evaluation of function of day to day activities nor do they present the patient's perspective of how he or she is functioning. The generic health status measures, by virtue of their use across different diagnostic groups, may have content irrelevant to the population under study and may lack evaluation of important components for the sarcoma population. As a result, the generic measures may be insensitive to important changes in status (Beaton, in press; Patrick, 1989).

Disease-specific measures similarly lack the content sufficient to evaluate the sarcoma population. The cancer-specific measures (e.g. EORTC-QLQ-C30, CARES) have a large proportion of content related to the side effects of chemotherapy with a small number of questions related to physical function (Aaronson, 1993; Ganz, 1994). A number of orthopaedic outcome questionnaires have been developed related to function at a specific joint such as the Lequesne Hip Score and the Simple Shoulder Test (Lequesne, 1980; Lippitt, 1993). However, a feature unique to reconstruction following sarcoma excision is the frequent involvement of the entire limb. Consequently, the existing measures lack characteristics that would adequately evaluate physical function and the effectiveness of treatment in the sarcoma population from the patient's perspective.

In addition to evaluating the effectiveness of treatment (including surgery with or without adjuvant therapy and rehabilitation) of sarcoma, there is a need to evaluate the disability level of these patients. Following surgery, rehabilitation is required to facilitate the patient's recovery. The philosophy of rehabilitation is to minimize disability and, when necessary, to provide adaptive methods or assistive devices to maximize abilities. However, those disabled as a result of their disease and its treatment may still require significant physical, economic and social supports. An instrument that is able to provide
an objective measure for evaluating the status of these individuals within the conceptual framework of the existing disability models (Nagi, 1965; Nagi, 1991; Wood, 1980; Verbrugge, 1991) would assist in classifying levels of disability. Such a measure would inform the rehabilitation model of disability evaluation and minimization. However, the usefulness of an outcome measure in evaluating treatment effects and describing an individual's disability is dependent upon the measurement properties, (i.e. reliability and validity), of the instrument.

In recognition of the need for an outcome measure to evaluate treatment effectiveness and describe disability levels in the lower extremity sarcoma population, we developed The Toronto Extremity Salvage Score (TESS). The TESS is intended to evaluate the domain of physical function in patients with bone and soft tissue sarcoma of the extremity. The items included in the questionnaires relate to activity limitations and restrictions in mobility, self-care, and in performing daily tasks and routines. The questionnaire is self-administered and reflects the patient's perception of his/her function. Preliminary developmental studies (presented in Chapter Two), item generation and reduction, test-retest reliability and known group validity were conducted for the measure as partial requirement for a Master's of Science (Davis, 1994; Davis, 1996). The goal of this thesis is to confirm the content and further validate the TESS, lower extremity version.

In summary, sarcoma is a life threatening disease but a significant number of patients (sixty to seventy percent) survive their disease and live with the resulting disability. A relevant outcome measure with proven measurement properties is required to describe the disability experienced by this population and to evaluate new treatment interventions. Chapter Two reviews the theoretical underpinnings and conceptual framework of the TESS in relation to the measurement principles of health status evaluation and the existing disability models. The results of the preliminary developmental
studies of the TESS are reviewed prior to presentation of the methods and the results of the current study.
CHAPTER TWO

FOUNDATIONS OF THE
TORONTO EXTREMITY SALVAGE SCORE

INTRODUCTION

The Toronto Extremity Salvage Score (TESS) is a measure of physical function intended to describe the disability experienced by the lower extremity sarcoma population and to evaluate the effectiveness of treatment interventions for this population. This chapter provides the background information for developing the TESS for these purposes.

THE TESS AS A MEASURE OF DISABILITY

Three major models of disability were developed over the past three decades: 1) that of Nagi (1965; 1991); 2) the International Classification of Impairments, Disabilities and Handicaps (ICIDH) of the World Health Organization (1980); and, 3) the model of Verbrugge (1994), which is an extension of the Nagi model. The disability models were conceptualized to create a set of inter-related concepts that would allow description and investigation of the consequences of disease beyond classification by disease entity (Nagi, 1965; WHO, 1980). Disease is described as progression from etiology to pathology that is manifested in signs and symptoms. Classifying by disease entity does not provide any information regarding the consequences of the disease as the severity and impact of the disease for the individual are not implicit in a diagnosis. The goal of the disability models
was to develop a framework that had a health context yet also provided a link to the larger problem of disability.

The development of a disability model was driven by rehabilitation, chronic disease interest groups and compensation and insurance companies (Albrecht, 1992; Nagi, 1991). The high prevalence of chronic illness and the moneys paid in compensation to injured workers have been instrumental in raising the question of the validity of disability claims that are based upon diagnosis and impairment descriptors. The questionable validity of diagnosis and impairment measures in determining disability arises due to the fact that severity is not an intrinsic component of all disease (e.g. different individuals all with osteoarthritis may have disease manifested over the spectrum of severity). While disability follows from impairments not all impairments result in disability (WHO, 1980). Badley (1987) has shown that there are low to moderate correlations between impairments and disabilities for individuals with stroke, multiple sclerosis, and rheumatoid arthritis and osteoarthritis.

The extremity sarcoma population is further representative of the problems of using impairment measures to quantify physical function. Clinical observation suggests that these patients rarely experience abnormalities at a single joint or in an isolated muscle group. For example, a patient with a soft tissue sarcoma of the thigh may experience hip and knee stiffness and weakness of muscles powering both joints. There may be oedema of the entire lower extremity. Consequently, the problem becomes one of rehabilitating the entire extremity with the goal of facilitating return of the individual to his or her maximal obtainable level of function. While the value of symptom control and impairment measures is recognized in monitoring clinical practice, the goal of rehabilitation is restoration of loss of functional ability (Albrecht, 1992) and thus, is best represented by the concept of disability.

Nagi (1965) was the first to formulate a disability model and the WHO (1980) and Verbrugge (1994) subsequently published models. The Nagi and Verbrugge models are
conceptual frameworks while the WHO model was formulated to create a taxonomic classification that was compatible with the International Classification of Disease (ICD). The authors of all three models agree that the disability literature traditionally has been plagued by a lack of conceptual clarity and inconsistent terminology across disciplines (Nagi, 1965; WHO, 1980; Verbrugge, 1994). The models are an attempt to provide conceptual clarity that will allow hypothesis testing and ultimately a better understanding of disability.

The ICIDH "provides a conceptual framework relevant to the long-term consequences of disease, injuries, disorders and is applicable to health care for early identification and prevention and to the mitigation of environmental and societal barriers" (WHO, 1980). The levels of the conceptual model consist of diseases, impairments, disabilities and handicaps. Nagi (1965) uses the terminology of active pathology, impairment, functional limitations and disability for his model and Verbrugge (1994) uses pathology, impairment, functional limitations and disability. These three models vary slightly in nomenclature and defining terms. Although each model has four conceptual levels, the comparative levels of the models do not mirror each other. For example, Verbrugge's description of functional limitations (Verbrugge, 1994) includes concepts that Nagi (1965) and the WHO (1980) include in impairments and Nagi's definition of functional limitations includes concepts found in both the disability and handicap definition of the ICIDH. Appendix One provides a comparative discussion of the models.

The ICIDH (WHO, 1980) is perhaps the most well known of the disability models. The ICIDH conceptualizes "diseases" as abnormalities occurring within an individual either at birth or acquired later in life. The underlying etiology of the disease may be sub-clinical or the disease may be manifested in signs and symptoms without recognition of the underlying pathology. The next level is "impairment" which is any loss or abnormality of psychological, physiological or anatomical structure or function. An example of an impairment in the musculoskeletal system would be loss of motion at a
joint. "Disability" is any restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being. These activities are represented by skills, tasks and behaviours. Disability theoretically results from an impairment but not all impairments result in disability. The final level of the ICIDH (WHO, 1980) is "handicaps". Handicap results from impairments and disabilities and is the limitation or prevention of a role that is normal (depending on age, gender, social and cultural factors) for the individual. Handicap includes six survival roles: orientation, physical independence, mobility, occupation, social integration and economic self-sufficiency, and any other concepts that might lead to disadvantage. There is an inherent value judgement in that handicap is judged by oneself and/or peers in relation to social and cultural norms. Nagi's model (Nagi, 1965; Nagi, 1991) presents problems for conceptualizing physical function as defined for the extremity sarcoma population. The comparable level to "disability" in the ICIDH (WHO, 1980) for Nagi (1965; Nagi, 1991) is "functional limitations". Functional limitations are the limitations "set on by impairments in an individual's ability to perform tasks and obligations of his usual roles and normal activities" (Nagi, 1965, p. 314). Consequently, these limitations are not only dependent upon the level of severity of the impairment but also on role demands and expectations. The final level of the Nagi model (Nagi, 1965) is "disability". Disabilities occur at the level of social functioning and are limitations or the inability to perform socially defined tasks and roles expected of an individual within a sociocultural and physical environment. These tasks and roles are subsumed in family/interpersonal relations; work, employment and other economic pursuits; education, recreation and self-care activities. The major differentiation between functional limitations and disability is that the former applies at the individual level whereas the latter is at the societal level (Jette, 1994). However, this differentiation becomes unclear by the use of "role" in both the functional limitations and disability categories. "Role" and "obligations" (used in functional limitations) imply an expected level of performance and the expectation level can be internally or externally
determined. In the disability category, however, the "roles" and "tasks" are explicitly stated by Nagi (1965) as those "expected within a sociocultural and physical environment". Does this then imply that the roles and obligations referred to under functional limitations are based only on the internal expectations of the individual? If so, Nagi needs to be more explicit in his definition of functional limitations. If Nagi meant functional limitations to refer to more elemental levels of purposeful activities at the individual level and disability to refer to the actions within the sociocultural and physical environment, inclusion of self-care within disability disputes this interpretation. The boundary between functional limitations and disability is unclear in the Nagi model. Furthermore, if Nagi's definitions (Nagi, 1965; Nagi, 1991) are compared to the ICIDH definitions of disability and handicap, Nagi's functional limitations, by virtue of the lack of definitional clarity, seem to include components represented by the ICIDH handicap.

Verbrugge's level comparable to the ICIDH "disability" is called "functional limitations". Functional limitations are defined as restrictions performing fundamental physical and mental actions (tasks) used in daily life by one's peers (Verbrugge, 1994). The actions are defined as generic tasks that might be used in many different situations. However, as a component of this category, Verbrugge (1994) includes many "tasks" such as "discrete motions and strengths, vision, hearing". This is contrary to Nagi's model (Nagi, 1965; Nagi, 1991) and the ICIDH (WHO, 1980) in which these items would be included within impairments. Consequently, the lower boundary of Verbrugge's (1994) functional limitations is shifted downward to overlap with impairment as defined by the Nagi (1965) and WHO (1980) models.

The definition of physical function applied to the extremity sarcoma population is not intended to extend to either the impairment level or a higher socially defined level of function. Consequently, the ICIDH (WHO, 1980) definition of disability seems to provide the most informative conceptualization of physical function for the extremity population. Physical function as defined by the TESS is subsumed in the ICIDH
definition of disability, limitations or restrictions in activities of the body as a whole. Physical function for the purposes of the TESS includes activity limitations such as restrictions in body movement (e.g. bending over), restrictions in mobility (e.g. walking, driving a car), restrictions in self-care (e.g. bathing) and restrictions in performing other daily tasks and routines (Patrick, 1993). The TESS, therefore, represents a measure of disability as defined by the International Classification of Impairments, Disabilities and Handicaps (WHO, 1980). Application of the taxonomic codes to the items of the questionnaire further confirm that the TESS is a disability measure (Appendix One).

OVERVIEW OF MEASUREMENT THEORY

Purpose of Measures

A number of authors (Feinstein, 1987; Guyatt, 1992; Kane, 1981; Kirschner, 1985; Williams, 1992) have defined the purposes of measurement and, although their terminology may be different, there is significant overlap in their definitions. Irrespective of the author, the three fundamental purposes of measurement include description at a single point in time, measurement of change in status (either in response to different treatments or over time), and prediction of a future event. The purposes of measurement are hierarchical as a measure that is capable of prediction also has properties such that it can evaluate change in status or status at a single point in time (Feinstein, 1987; Kane, 1981; Williams, 1992).

Measures may be applied at a single point in time to describe a group or an individual. Such measures may be used for the purpose of developing normative data and, because these measures are taken only once, test-retest reliability and detection of change are not necessary. Rather internal consistency and face and content validity are requisite properties.

Measures may also be taken at a single point in time for the purpose of screening. The goal in this setting is to identify individuals at high risk who require further
evaluation. Therefore, sensitivity and specificity are important. The priority is to maximize sensitivity.

The next level of measurement is the evaluation of change in status, either in response to different treatments or over time (Feinstein, 1987; Kane, 1981; Williams, 1992). Measurement error is important in determining true change versus change as a result of an unreliable measure. Test-retest reliability is, therefore, important to evaluate as is sensitivity to change. Change may be quantified and evaluated statistically, however, it is also important to know how much change is clinically meaningful.

The most stringent level of measurement is prediction or prognostication. These measures are applied to predict future events and, thus, require serial measures over time. The measure must be able to determine change so test-retest reliability is again important. A further criteria, that of predictive validity, is required as the measure must have strong relationships to the clinical course and occurrence of a future event (Kane, 1981). In reviewing the properties required for a predictive measure, the hierarchical nature of measurement purposes is noted. The properties required for prediction subsume those required for evaluating change in status and describing status at a single point in time.

**Group Versus Individual Measurement**

Measures may be used with groups or individuals, irrespective of the purpose for which a measure was developed. However, the standards for measures of groups versus individuals are different with group criteria having lower standards (Long, 1996; Nunnally, 1978; McDowell, 1996; McHorney, 1995). A reliability coefficient of 0.80-0.90 is acceptable when reporting data for groups (Nunnally, 1978) whereas a coefficient in the range of 0.90 to 0.95 is recommended for making decisions about individuals (McHorney, 1995; Nunnally, 1978; Streiner, 1995). Fleiss (1986) has shown that low levels of reliability significantly increase sample size requirements in clinical trials. The
recommended higher reliability coefficient required for individual patient monitoring indicates improved precision of the measure, i.e. it has reduced error of measurement.

The TESS requires measurement properties for evaluating treatment effectiveness between groups of patients. However, the other goal of developing the TESS is to be able to monitor patients over time and to use the information to educate patients regarding their anticipated recovery. This requires that the TESS have reliability and validity such that the scores on the measure accurately reflect the clinical course of individual patients. Consequently, both group and individual standards of measurement, i.e. reliability coefficients in the range of 0.90 to 0.95 (McHorney, 1995), must be met for the TESS to adequately serve its intended purposes.

Types of Measures

Patrick (1993) notes that health-related quality of life or health status questionnaires may be generic, disease-, region- or domain-specific. The generic measures have been designed for use irrespective of diagnosis and consequently may not contain items that are specific to the disease population being evaluated. As a consequence, researchers have concluded that generic measures are unlikely to be sensitive to changes in status (Patrick, 1993). These generic measures, however, have the advantage of allowing comparison of outcomes among different diagnostic groups, especially as the measurement properties of a generic measure are evaluated in multiple disease populations. National normative data are also available for some measures e.g. the SF-36 has normative data for England, France, Germany, Sweden, Wales and the United States. Disease-specific measures are designed for use with specific diagnostic groups and contain items specifically relevant to the targeted disease e.g. the WOMAC for osteoarthritis of the hip and knee (Bellamy, 1988). A disease-specific measure is anticipated to be more sensitive to change in status but has the disadvantage of allowing comparison of outcome only within similar diagnostic groups. Region-specific measures are designed for use in particular anatomical
locations. For example, there are a number of measures that have been designed for use in the shoulder (Constant, 1987; Lippitt, 1993; Roach, 1991). System-specific measures e.g. gastrointestinal, have also been developed (Eyspach, 1993).

Whether generic, disease-, region- or system-specific, most measures evaluate the three primary domains of function, that is, physical, emotional and social function. There are, however, situations where domain-specific measures are appropriate and this traditionally has been most prominent in mental health e.g. Beck depression scale, CES-D (Beck, 1961; Radloff, 1977). These scales evaluate only the domain of mental functioning.

The TESS is a domain-specific measure in that it is intended to evaluate only physical function. It is also region-specific in that the measure evaluates physical function of the lower extremity. It was conceptualized for the extremity sarcoma population and developmental studies to date have used the TESS only with the extremity sarcoma population, suggesting that the measure is disease-specific. However, Amadeo (1996), in his use of a carpal tunnel measure for evaluating Colle’s fracture patients, has shown that a disease-specific measure can provide meaningful information when used in conditions other than that for which it was designed. The TESS is currently being used with total hip revision and hip fracture populations (Gross, personal communication; Jaglal, personal communication). These populations all involve lower extremity pathology and it remains to be seen whether the TESS is disease- or region-specific.

Methods of Measure Development

Rating scales measuring clinical phenomenon can be developed using clinimetric (Feinstein, 1987) or psychometric approaches (Nunnally, 1978). Irrespective of their method of development, all measures must be reliable and valid. Reliability and validity may be assessed in different ways depending on whether the focus has been on clinimetric or psychometric methods.
The clinimetric approach relies on clinical judgement to develop measures that are useful in evaluating therapeutic effectiveness and consequently, often lacks a formal methodological or conceptual framework (Feinstein, 1987; Guyatt, 1986). Items are generated based on clinician experience and/or patients' perceptions of what is important. Ratings of relevance or importance form the basis for item inclusion in a measure. As the items are selected for their unique contribution rather than their measurement properties, heterogeneity is often an aim of the measure (Wright, 1992). The purpose of clinimetric measures is to evaluate clinical effectiveness, therefore, clinical sensibility and responsiveness are important criteria for judging the performance of the measure.

Psychometrically developed measures traditionally originated in the social and behavioural sciences. The methods were used to select and aggregate items into measures of theoretical constructs. Psychometrically developed measures have usually been applied in fields lacking external criteria for evaluating validity (e.g., intelligence testing). Items for the measure are chosen based on the consistency of the responses and the degree to which the pattern of responses conforms to the underlying construct. Statistical methods of internal consistency and factor analysis are usually used in determining the final content of the measure. The construct validity (convergent and discriminant) of the measure is evaluated by hypothesis testing (Kerlinger, 1986; Nunnally, 1978).

While clinimetric and psychometric approaches use different methods, they share common phases of development: item generation; item reduction; and, aggregation to summary scores. However, psychometric methods are becoming more common in the development of clinical measures, particularly in the development of health status measures (e.g., Sickness Impact Profile, SF-36, Functional Independence Measure), where reliability and validity testing are now commonly included in the development of clinical measures. Consequently, clinimetric and psychometric methods may be considered as complimentary approaches to measure development (Davis, submitted).
In clinimetrically developed measures, Feinstein (1987) uses the term "sensibility" to include the purpose of the measure, framework, overt format, face validity, content validity and ease of use of a measure. The evaluation of sensibility relies on qualitative judgements using clinical knowledge. In looking at the measure, one judges whether the items are understandable and appear to make sense in relation to the clinical problem. In contrast to sensibility, the evaluation of reliability and validity requires quantitative evaluation using statistical methods.

Internal consistency is frequently assessed in psychometrically developed measures. This type of reliability evaluates how subjective responses to an item by participants relate to each other and how each item relates to the total score. Each item is supposed to be related to the underlying construct of the scale and, therefore, the summing of the items provides more information than the individual items. There is, however, no external measure by which to demonstrate the relationship of the items to the construct. Cronbach's alpha (1951) is a measure of the strength of internal consistency. Alpha is directly related to the number of items in the measure and the mean item-to-item correlation. The greater the number of items and the higher the mean item-to-item correlation, the higher the alpha coefficient. Nunnally (1978) recommends that for a measure to be used in clinical research to assess important clinical differences between groups, the reliability coefficient should be 0.90 or higher and at least 0.95 if the measure is to be used in making treatment decisions regarding individuals (Nunnally, 1978). A reliability coefficient of 0.80 is acceptable for the first stages in the development of a measure.

Reliability also refers to the consistency of the results obtained under stable patient conditions. In this situation, there is an external criteria (e.g. stable patient status) against which to judge the ratings. Test-retest reliability is evaluated in the case of self-administered measures with intra- and inter-rater reliability evaluated in interviewer administered measures. Reliability is the proportion of variance attributed to the real score
differences between subjects and should be large compared to error. The intraclass correlation coefficient (ICC) for continuous data (Fleiss, 1986), kappa for categorical data (Landis, 1977) and weighted kappa (Landis, 1977) and Kendall's coefficient of concordance (Sprent, 1993) for ordinal data are statistical tests of concordance. In the case of the fixed effects model (in which the raters are the only ones who will use the measure), the ICC and Cronbach's alpha are equivalent (Bravo, 1991). As previously noted, high levels of reliability (i.e. > 0.90) are required for making individual treatment decisions while lesser levels (i.e. > 0.80) are acceptable for groups. While group evaluations may allow for lower reliability coefficients, these coefficients reflect less precision in the score and consequently limit the power to detect small effect sizes without significant increases in sample size (Fleiss, 1986; Streiner, 1995).

Validity of a measure is required to determine if the instrument is measuring what it says it does. Reliability is a necessary criterion for validity but it does not ensure that the measure is valid. Face and content validity are evaluated qualitatively. However, there are also empirical measures of validity. These include evaluating the depth of a measure or its ability to evaluate the spectrum of the severity of a construct by calculating the prevalence of floor and ceiling effects (McHorney, 1995). Floor and ceiling effects occur when the distribution of scores on a measure has a large proportion of very low scores or alternately a large proportion of high scores. McHorney (1995) suggests that floor and/or ceiling effects of greater than 15% are unacceptable in a measure. Criterion and construct validity are also evaluated quantitatively and are established by the testing of a priori hypotheses.

Criterion validity is evaluated by comparing a measure to an existing external criterion of the phenomenon under study (Last, 1995). There are two types of criterion validity: concurrent validity and predictive validity. Concurrent validity evaluates the relationship of the new measure with the current standard criterion measure at a single point in time. A strong positive correlation between the two measures would provide evidence of concurrent validity. If the result of the relationship between the measure and
the criterion event is known only at a future date, it is called predictive validity (Feinstein, 1987; Last, 1995). Spector (1992) further described a sub-category of concurrent validity, "known group validity". Known group validity uses as a criterion variable a categorical measure that sub-divides the sample into different levels. The relationship of the new scale compared to the known group criterion is then evaluated.

A further criterion necessary for measures used in monitoring patients is the ability of the measure to detect change. In the clinical setting, the ability of a measure to detect change has been called responsiveness (Guyatt, 1987; Medical Outcomes Trust Source Pages, 1996). This may be change within a subject over time or, when used as an outcome in a clinical trial, may be the difference between treatment groups. Effect sizes (Cohen, 1988; Kazis, 1989; Liang, 1990) are standardized measures of change and are most commonly used as a measure of responsiveness. Effect sizes have been calculated using different methods by different authors; the difference is divided by a form of the variation. Cohen (1988) uses the standard deviation of the differences and with paired data considers the correlation of the scores at the two times. The method of Kazis (1989) uses the baseline standard deviation. Guyatt's index of responsiveness (1987) uses the variation of a stable control group. The method of Cohen (1988) probably most accurately reflects responsiveness as it incorporates the baseline heterogeneity and variability of response to treatment. The use of only the baseline standard deviation (Kazis, 1989) produces effect sizes dependent on the baseline heterogeneity of the sample while the index of responsiveness (Guyatt, 1987) likely under estimates the variability of the sample by evaluating change in relation to the variability in a stable control group. Consequently, the method of Cohen (1988) is preferred in evaluating the responsiveness of health status measures.

In the special case where there is a treatment of known efficacy, Liang (1985) describes the relative efficiency of two measures as a means of evaluating responsiveness.
Using a ratio of t-tests where sample sizes are constant, a more efficient measure in terms of sample size requirements and study power has a ratio greater than one.

Construct validity differs from criterion validity in that it tests the underlying conceptual framework of the scale (Feinstein, 1987). Construct validity is most commonly evaluated by testing a priori hypotheses about how the construct relates to other constructs. The strength and direction of the anticipated relationship is stated. Convergent validity is tested by seeing how strongly the construct correlates with similar constructs to which it should be related. Alternately, divergent or discriminant validity is demonstrated when constructs that should be unrelated are poorly correlated.

In summary, reliability and validity are measurement properties that must be evaluated in determining the usefulness of a measure. The standards required for each of these properties are based on qualitative and quantitative methods and require clinical judgement. The necessary properties depend on the hierarchical purposes of measures and the standards of the properties are dependent on whether the measures will be used for groups or individuals.

EXISTING MEASURES OF HEALTH STATUS USED IN CANCER POPULATIONS

Numerous books review the existing measures available for evaluating health status and functional outcomes in various populations (Bowling, 1991; Bowling, 1995; Cole, 1994; McDowell, 1987; McDowell, 1995; Osoba, 1991; Patrick, 1993; Spilker, 1996). The purpose of this section is to provide an overview of commonly used measures that might be relevant to the extremity sarcoma population. These include the existing generic measures (e.g. Sickness Impact Profile, Nottingham Health Profile and the SF-36), the cancer-specific measures (e.g. Karnofsky Performance Status measure, EORTC-QLQ-C30 and CARES), and the disease-specific measures (Musculoskeletal Tumor Society Rating Scale, 1987 and 1993 versions).
While it is recognized that this review of existing measures is not exhaustive, it highlights the issues with the existing measures that justify the development of a new measure for the extremity sarcoma population.

**Generic Measures**

Generic health status measures have been used across types and severity of diseases, treatment and patient populations (Patrick, 1989; Patrick, 1993) and their content, at minimum, evaluates the domains of social, emotional and physical functioning. Generic measures have traditionally been used for group comparisons of health profiles of different diagnostic groups (Jenkinson, 1988; Stewart, 1989), evaluation of treatment interventions between groups (Bombardier, 1986) and for describing population health (Hunt, 1984). The Sickness Impact Profile (Bergner, 1976; Bergner, 1981), Nottingham Health Profile (Hunt, 1981; Hunt, 1985) and the SF-36 (Ware, 1992; Ware, 1993) are three well known generic measures.

The Sickness Impact Profile (SIP) is a multi-dimensional measure that reflects the effect of illness on daily activities, feelings and attitudes as perceived by the respondent (McDowell, 1987). The items were gathered from extensive surveys of health professionals, patients, care providers, and healthy individuals (Bergner, 1976a). Items were retained for their uniqueness of content, and were then subsequently reduced using mathematical processes, mainly regression methods (Bergner, 1976a). The instrument has 136 items of which forty-five are included in physical function. This domain includes items related to mobility, ambulation and body movement. Upper extremity activities are poorly represented by only four items and a further five items are not applicable to the extremity sarcoma population (e.g. items related to bowel and bladder control). The response options for the SIP are dichotomous which limits the ability of the instrument to detect gradations of change within the items. The reliability and validity of the SIP has been demonstrated in many chronic disease populations (Bergner, 1976b; Bergner, 1981;
Deyo, 1983; Jones, 1989). Patrick (1993) summarizes the SIP scores from twenty-seven different studies of adults with various medical conditions. These data suggest that the SIP is able to appropriately discriminate among medical populations of varied severity. The responsiveness studies of the SIP have produced variable results and seem to suggest that the measure is more sensitive to degradation in function (Deyo, 1984; MacKenzie, 1986).

Sugarbaker (1982) used the SIP as one of over twenty outcomes (a primary end point was not identified) in a study randomizing twenty-six soft tissue sarcoma patients to treatment with limb-sparing surgery, irradiation and chemotherapy, or to treatment with amputation and chemotherapy. Two dimensions of the SIP, emotional behaviour and body care and movement, showed statistically significant differences between the two treatment groups on a single post-operative measure. The significant results represent a minimal component of a profile of outcomes and it is impossible to interpret these data within the context of the profile of outcomes.

The Nottingham Health Profile (NHP) stemmed from the Sickness Impact Profile and measures perceived physical, social and emotional health problems (McDowell, 1987). It was initially developed to predict the need for health care services from health surveys in the British population. The items of the NHP were derived from interviews with patients suffering from various acute and chronic conditions. After eliminating redundant and ambiguous statements, thirty-eight items were retained. There are two parts to the NHP, the first of which evaluates six domains of distress and the second of which evaluates the impact of health problems on seven components of everyday life. Responses are rated dichotomously and each domain is given a summated score. Thurston scaling methods were used to derive item weights reflecting gradation in severity (McKenna, 1981), however, the validity of the weights has been questioned as the physical function domain inappropriately orders individuals along the severity continuum (Wilkin, 1992). The authors insist that the NHP scores be analyzed with nonparametric statistics. The NHP is most useful in the evaluation of chronic, disabling illnesses in the elderly.
population (McEwen, 1996). The dichotomous responses of the NHP and the item content which reflects its initial purpose, the need for health care services, limit its relevance to the sarcoma population.

The SF-36 is another generic health status instrument that has been marketed worldwide (Aaronson, 1992) and the use of this measure in clinical trials and individual patient monitoring (Ware personal communication, 1996) is highly recommended. The current study provided the opportunity to evaluate the SF-36 as an outcome measure for the extremity sarcoma population.

The SF-36 was developed from the Health Insurance Experiments (HIE) and the Medical Outcome Studies (MOS) with the intent of representing physical and mental health concepts based on the inclusions of behavioural function, perceived well-being, social and role disability, and personal perceptions of health in general (Ware, 1993). The initial questionnaire was the SF-20 with eighteen items drawn from the HIE and MOS and two single items, social functioning and bodily pain, added (Stewart, 1988). However, low levels of internal consistency (0.76) and floor effects that occurred in seriously ill patients (Bindman, 1990) resulted in the development of the SF-36. A detailed summary and critique of the development of the SF-36 presented in Appendix Two is summarized below.

The SF-36 evaluates the domains of physical function, role limitations due to physical function, bodily pain, vitality, general health, role limitations due to emotional problems, social function, mental health and change in health. These eight health concepts (Ware, 1992) are representative of the minimum standards required for a comprehensive health status measurement with the addition of the concepts of bodily pain and vitality. Health transition is an additional concept.

Ratings of the SF-36 items are made on a Likert-type response scale based on a four week or one week recall depending on the questionnaire version used. With the exception of the health transition item, each of the subscales provides a standardized score
ranging from zero to one hundred. Ware (1992) justifies the assignment of an item to a specific subscale based on the item's correlation with the overall subscale score, that is, the item has the highest correlation with the subscale of which it is a member. However, in reviewing the correlations (Ware, 1992), a number of items also have moderate correlations (range 0.40 to 0.60) with multiple subscales suggesting that item-to-subscale relationships are more complex and inter-related than suggested by Ware (Appendix Two).

Reliability, validity and responsiveness for the subscales have been evaluated in multiple chronic and acute disease populations (Jenkinson, 1996; Ware, 1993). Overall, the results of these studies suggest that the SF-36 subscales have reliability sufficient for group comparisons with \( r=0.78-0.93 \) with 95% confidence intervals of the standard error of the mean ranging between thirteen and thirty-two in cross-sectional studies and test-retest reliability coefficients of 0.60 to 0.81 with 95% confidence intervals around the standard error of the mean of nineteen to forty-seven (McHorney, 1995). Validity (Aaronson, 1992; Brazier, 1992; McHorney, 1993; Ware, 1992; Ware 1993) of the SF-36 subscales has been demonstrated against clinically defined criterion groups and long-form versions of the MOS scales of known validity. These analyses have shown that the subscales successfully distinguish psychiatric and physical illnesses, and discriminate severely ill medical groups from moderately ill and healthy groups for some subscales (Ware, 1993). The physical and mental health subscales appropriately discriminate moderate levels of illness burden in patient groups with physical or psychiatric illness. However, the bodily pain scale had poor convergent validity with groups of known disease severity (McHorney, 1993). Floor effects have been noted in role functioning in severely ill populations, specifically 24% for role-physical and 18% for role-emotional (Kurtin, 1992; McHorney, 1994). Complete review of the developmental process, psychometric properties and normative data for England, Wales, France, Germany, Sweden and the United States have been published for the subscales. The intent of the SF-
36 was to meet the minimum criteria for group evaluation and the subscales generally meet these criteria.

Ware (1994) subsequently published data in support of an SF-36 Physical and Mental Health summary scale. Ware (1994) suggests that there is a two factor solution for the SF-36, physical and mental health, that accounts for eighty to eighty-five percent of the variance of the eight subscales. It was hypothesized that the physical function, role-physical and bodily pain subscales would have the highest correlations with the physical health factor and that the social function, role-emotional and mental health subscales would have the highest correlations with the mental health factor. General health and vitality were anticipated to correlate with both the physical and mental health factors. Two factors with eigenvalues greater than one were extracted after orthogonal rotation and as hypothesized the subscales were correlated with the two factors. (An explanation of factor analysis is presented in the following methods chapter as this technique was used in analyzing data for this thesis.) Ware's interpretation of these data is questionable, however, as simple structure has not been obtained through orthogonal rotation.

The goal of principal component analysis is to obtain simple structure, that is the least number of factors to explain the variance with each item (subscale in this case) correlated greater than 0.35 (Nunnally, 1978) with only a single factor. The orthogonal solution presented by Ware (1994) has three subscales, general health, vitality and social function loading on both factors and this does not achieve simple structure or support the two factor solution for the SF-36.

Other authors (Jenkinson, 1996; Wolinsky, 1996) have also factor analyzed the SF-36. Jenkinson (1996) analyzed the British data and obtained results similar to Ware (1995) with a "two factor" solution that accounted for sixty-six percent of the variability. Again, the subscales of general health, vitality and social function loaded on both factors. Of note in Jenkinson's results (1996) is the eigenvalue of the physical factor (1.17). Using an eigenvalue criteria of greater than or equal to one for maintaining factors, this
barely meets the standard for maintaining factors. The physical health factor accounts for a minimal amount of the variance. Wolinsky (1996) factor analyzed the thirty-six items of the SF-36 with the result that the data were best represented by nine subscales. The ninth subscale, labeled "health optimism" by Wolinsky (1996), included the items "getting sick" and "getting worse" as their own subscale. Despite Ware's interpretation of his factor analysis (Ware, 1995), the results of the above studies question the validity of the SF-36 physical and mental health summary scales. Wolinsky's (1996) results also question the validity of the eight subscales defined by Ware (1993).

The generic health status measures were traditionally developed for use in ambulatory, community dwelling populations and their content, therefore, reflects the broad scope of health experience of the population. The SIP, NHP and SF-36 have been used in many populations and there are some variations in their performance with different disease groups. Given the broad scope of the measures this is not perhaps surprising and suggests that there may be certain disease groups where one generic measure may have superior performance to another. Beaton et al (1997) compared the reliability and responsiveness of the pain and physical function subscales, and the overall score of five generic health status measures in injured workers. These authors found that the SF-36 pain and physical function subscales and overall score had adequate levels of test-retest reliability for groups and was superior on several statistics of responsiveness when compared to the similar scores of other generic measures (including the NHP and the SIP). However, the criteria used to classify test-retest reliability as "adequate" was based on intraclass correlation coefficients ranging from 0.74 to 0.85 (Beaton, 1997) and this range is below the 0.85 to 0.90 recommended by Nunnally (1978) for group evaluation. Even though the SF-36 scales were more responsive than those of the other generic measures, the low level of test-retest reliability suggests that the SF-36 scores had limited precision, that is, a large error component. The selection of a generic measure requires
review of the sensibility and psychometric properties of a given instrument prior to its use with a given population.

McHorney (1995) evaluated several health status measures based on criteria for use in making clinical decisions for individuals. In relation to group versus individual subject measurement, the cross-sectional and longitudinal precision of the instruments were evaluated. The recommended criterion is 0.90 to 0.95 for individuals when using internal consistency as a measure of cross-sectional precision. Only the physical function and mental health subscales of the SF-36 met the lower criteria of 0.90 with none of the remaining six SF-36 subscales, the NHP, the Functional Status Questionnaire (FSQ), the Dartmouth COOP Poster Charts (COOP) or the Duke Health Profile (DHP) meeting the standards for individual patient use (Hunt, 1980; Jette, 1986a; Nelson, 1987; Parkerson, 1990; Ware, 1993). Similar results were found in evaluating longitudinal precision (test-retest reliability) where the scales of all the measures were below the 0.90 level (the lower bound for individual patient use) with wide confidence intervals. None of the measures evaluated met the standards for use with individual patients and this is despite the FSQ and the COOP being developed for individual patient use (Jette, 1986a; Nelson, 1987). The intent of the SF-36 was to meet the minimal criteria for use with groups, not to achieve the higher standards necessary for individual use (Ware, 1993). However, Ware (1996) continues to present examples and interpretation of SF-36 data for individual patients. Users of measures need to be aware of the adequacy of the various measures in evaluating groups versus individuals such that they make informed decisions about their chosen measures.

**Cancer-specific Measures**

Cancer-specific measures of health status date back to the development of the Karnofsky Performance Status (KPS) scale (Karnofsky, 1948; Karnofsky, 1949). In the intervening years, a number of measures have been developed but the emphasis of the
instruments has frequently been related to the side-effects of chemotherapeutic agents and their immediate effect on functional status. The instruments developed in more recent times have adopted the inclusion of symptoms, social, emotional, and physical functioning in evaluating health status. Numerous books have been written detailing measures for the cancer population (Bowling, 1995; Osoba, 1991; Spilker, 1996) and only those measures most commonly used are detailed here. The KPS, the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30) and the Cancer Rehabilitation System (CARES) are examples of cancer-specific measures currently used in evaluating health status for this population (Aaronson, 1988; Aaronson, 1993; Ganz, 1992; Schag, 1983; Schag, 1990a; Schag, 1990b).

The KPS (Karnofsky, 1948; Karnofsky, 1949) is a clinimetric measure of physical function for cancer patients; the content was determined solely on the clinical judgements of what needed to be included in the measure. It was initially used to evaluate the change in performance status of patients with bronchogenic carcinoma following treatment with nitrogen mustards (Karnofsky, 1948). An observer rates the subject's ability to continue normal activities, or his/her degree of dependence on assistance or nursing care. The KPS is measured in increments of ten ranging from one hundred (normal function with no evidence of disease) to zero (deceased). In relation to the extremity sarcoma population, the scoring levels of the KPS are very coarse (for example: 70=Cares for self. Unable to carry on normal activity or to do active work; 80=Normal activity with effort; some signs or symptoms of disease) such that discrimination based on disability level and the ability to detect change would be unlikely. Although the KPS is unlikely to be of use in evaluating physical disability in patients treated for extremity sarcoma, it is the most widely used and cited measure in cancer clinical trials, particularly for establishing subject entry criteria (Hutchinson, 1979; Spilker, 1992).
The QLQ-C30 is a cancer-specific measure developed specifically for use in clinical trials (Aaronson, 1988; Aaronson, 1993). The instrument is based on a modular concept in which the core instrument, the QLQ-C30, is used with a site-specific module e.g. lung cancer. The items of the QLQ-C30 were generated by an international panel of clinical experts and the initial version of the questionnaire consisted of thirty-six items that were reduced to thirty items after initial testing (Aaronson, 1991). The instrument consists of five functional scales: physical; role; emotional; social; and cognitive functioning; three symptom scales: fatigue; nausea / vomiting; and pain; a global health-related quality of life scale; and single symptom items of dyspnoea; sleep disturbance; constipation/diarrhoea; and perceived financial impact. Ratings for all items are based on the previous week and all responses are on a four-point Likert-type response with the exception of the physical and role functioning scales which use dichotomous responses. All scale responses are converted to range from zero to one hundred with higher scores being indicative of higher levels of function. The symptom items are similarly converted for scaling except that higher scores indicate greater symptomatology.

The initial psychometric testing of the QLQ-C30 was undertaken in a twelve country sample of lung cancer patients. Internal consistency reliability ranged from 0.52 to 0.89, with all but one of the nine multi-item scales having an alpha greater than 0.70. Test-retest reliability is adequate for group comparisons (Aaronson, 1993). Discriminant validity was demonstrated against groups of varying levels of performance status and known group validity has also been demonstrated (Aaronson, 1993). Responsiveness was demonstrated by statistically significant different changes in scores in the expected direction based on changes in performance status and clinical response to treatment. Further testing of the QLQ-C30 is being undertaken in other cancer sites e.g. head and neck cancer (Jones, 1992).

Irrespective of the psychometric properties of the QLQ-C30, there are fundamental problems with the measure in considering its use with the extremity sarcoma population.
Over one-third of the items deal with symptoms, a significant number of which are not applicable to the extremity sarcoma population e.g. nausea and vomiting, constipation and diarrhoea. The physical function and role items are multi-faceted in that a single item asks about difficulty in up to four items e.g. going upstairs and sitting on a toilet are included in a single item with two other activities. Consequently, the response is unclear as to whether it refers to difficulty with all or a portion of the activities. Perhaps of major concern is the dichotomous scaling of the physical function items. While scaling with more categories, usually five to seven levels, is no guarantee of responsiveness, multiple categories allow for a greater potential for the patient to express change than dichotomous response options. These characteristics of the QLQ-C30 result in the questionnaire having poor sensibility for the extremity sarcoma population.

Schag and colleagues have developed the CARES as a generic measure of health-related quality of life suitable for use in cancer patients (Ganz, 1992). The conceptualization of the instrument as generic is based on the perspective that cancer is a general diagnosis in which the natural history and treatment is dictated by the site and subtype of the disease. The content of the scale was derived from patients and health professionals. The CARES is a self-completed questionnaire and is rated on a five-point scale ranging from "does not apply" to "applies very much" based on the experience of the previous month. There is a global summary score and five higher order factor scales that are scored as subscales (physical, psychological, medical interaction, marital and sexual) and thirty-one more specific subscales. Depending on the applicability of items, scoring is based on a minimum of ninety-three ranging to a maximum of 132 items. The CARES has been evaluated in many cancer sub-types, including breast, colon, lung, prostate, and has demonstrated psychometric properties of reliability (adequate for group comparisons), validity and responsiveness (Ganz, 1992; Schag, 1983; Schag, 1990a; Schag, 1990b).

The CARES is inclusive of perceptions of interactions with the treatment team (e.g. nurses don't explain what they do), reaction to treatments (e.g. diagnostic
procedures painful), other patients (e.g. uncomfortable to see patients get treatments), compliance and many other concepts that make it a very comprehensive but lengthy instrument. While the value of evaluating these concepts within a given context and for a specific purpose is recognized, a significant proportion of the content of the CARES is not applicable in the monitoring of patients with extremity sarcoma over time or in response to different treatment interventions.

**Disease-specific Measures**

The Musculoskeletal Tumor Society Rating Scale (MSTS) is the only disease-specific measure for the extremity sarcoma patient (Enneking, 1987). It was developed by clinicians in the early 1980's as a means of documenting surgical outcome and, after "field testing", the MSTS (Enneking, 1987) was revised to a seven item scale that is completed by the clinician. (The methods and results of the field testing process have never been published.) The seven items scored are: pain; range of motion; strength; joint stability; deformity; overall functional ability; and, emotional acceptance of the surgery. The extremity has been divided into regions: shoulder girdle/shoulder; elbow; forearm/wrist/hand; pelvis/hip; knee; and, foot/ankle. The seven items are rated according to criteria specific to each anatomical region. Five points are rated as excellent, three points are given a good rating, a fair is one point and no points are given for poor for each of the items for a maximum score of thirty-five points. An overall rating of poor, fair, good or excellent is also given based on the number of items receiving a rating of excellent, good, fair, or poor. If any two items receive a lower category rating, the overall categorical rating assigned is that lower rating. For example, five excellents, one good and one fair would receive an overall categorical rating of good; five excellents and two fairs would be an overall rating of fair; and, five excellents, one good and one fair would be an overall rating of good. This categorical rating is problematic, however, as it does not correspond to the numerical scores. As an example, a subject with seven goods would have a score of
twenty-one (7 x 3) and an overall rating of good. A subject with five excellents and two fairs would have a score of twenty-seven [(5 x 5) + (2 x 1)], higher numerically than the previous subject, but would be rated as fair, lower than the previous subject.

The content of the MSTS (Enneking, 1987) is heavily weighted with items reflecting impairment (range of motion, strength, deformity, stability) that do not necessarily reflect an individual's ability to complete his/her daily tasks and roles. The clinician rates the patient on the items and the literature would suggest that clinician and patient ratings differ significantly such that the scores on the MSTS may be a poor reflection of patient functional status (Meenan, 1982). The measurement properties of the MSTS are unknown as it has not been tested for reliability or validity.

The MSTS was revised in 1993 (Enneking, 1993) based on the recommendations of a working group of musculoskeletal surgical oncologists. This revision was in response to criticism of the previous version (Enneking, 1987), particularly with regard to the emphasis on impairment items and the scoring algorithm. The revised version of the MSTS (Enneking, 1993) consists of an upper extremity scale and a lower extremity scale each comprised of six items rated on a five point Likert-type scale. The upper extremity items are: pain; function; acceptance; hand positioning; dexterity; and, lifting ability. The lower extremity items are: pain; function; acceptance; use of ambulatory aids; walking ability; and a final item called "gait handicap". Items are rated by the clinician and may be coded as "not applicable". These "not applicable" items are eliminated in calculating the final score. The final score is a percentage. The psychometric properties of the scale have not been tested although Enneking (1993) reports that there is high percent agreement between the scores on the 1987 and 1993 versions of the MSTS. This revised version of the MSTS reflects many of the same problems of the older version, particularly regarding face validity, completion by the clinician and untested measurement properties.

The existing generic questionnaires, while providing the benefit of comparison of different diagnostic groups, do not alone provide sufficient information about the health-
related quality of life of specific diagnostic groups. The current recommendations are that health-related quality of life be evaluated using both generic and disease-specific measures (Guyatt, 1993; Nayfield, 1992; Patrick, 1989; Patrick, 1990; Sprangers, 1993). The existing cancer-specific instruments have tended to focus on the immediate effects of treatment interventions and consequently contain a large number of symptom and toxicity related questions. The MSTS (1987 and 1993 versions) has attempted to evaluate the extremity population but the content of each version of the measure provides a very limited definition of physical function and only from the clinician's perspective. Consequently, the existing cancer-specific (Aaronson, 1993; Schag, 1990a; Schag, 1990b) and sarcoma-specific instruments (Enneking, 1987; Enneking, 1993) conceptually and from measurement property criteria are inadequate to evaluate physical function for the extremity sarcoma population. The limitations of the existing generic, cancer-specific and disease-specific instruments provided the impetus for development of a new measure of physical function for the extremity sarcoma population.

DEVELOPMENT AND PRELIMINARY EVALUATION OF THE MEASUREMENT PROPERTIES OF THE TESS: LOWER EXTREMITY VERSION

The purpose of the TESS is to monitor the physical function of patients over time and to evaluate the effectiveness of treatment interventions both in clinical and research settings. The TESS reflects the patient's perception of his/her functioning over the previous week and patients need approximately ten minutes to complete the questionnaire. To date the TESS has undergone item generation and reduction, evaluation of item weighting, reliability testing and preliminary validity testing (Davis, 1994; Davis, 1996).

Items in the initial version of the TESS were based on a literature review and suggestions from the multi-disciplinary team members (surgeons, nurses, physiotherapists and occupational therapists) who treat the sarcoma patients. Patients were then mailed the lower extremity questionnaire and asked to rate the difficulty experienced
performing each activity on a five-point Likert type scale. They also rated the importance of each item on a four-point scale. Sixty-six lower extremity subjects (of seventy-five surveyed) who had completed treatment a minimum of six months previously responded.

A clinimetric strategy to item generation and reduction was used. Patients' perceptions of the importance and relevance of items based on the importance ratings and the frequency of "not applicable" responses were used to delete items. *A priori* criteria of thirty percent of ratings as "totally unimportant" or "not applicable" did not result in elimination of any items. Items omitted from this first version of the questionnaire were identified by allowing subjects to add up to five additional items that they felt were relevant to physical function. Using the thirty percent *a priori* criteria, one item, sporting activities, was added to the lower extremity questionnaire. The resultant lower extremity questionnaire has thirty items (Appendix Three).

The inclusion of a "not applicable" response option has implications for the creation of a summary score for the TESS. The "not applicable" items are excluded from the numerator and denominator of the TESS summary score. This makes the TESS patient-specific as it is based only on items that are applicable and important to the individual. This allows inclusion of a broad spectrum of disability items that are relevant to individuals of varying ages. For example, the extremity sarcoma population treated in an adult centre ranges from thirteen to eighty year olds and not all items are relevant to all subjects e.g. driving, gardening, sports. However, the inclusion of the items in the measure contributes to the face and content validity of the instrument for this heterogeneous population.

The trade-off in using a "not applicable" response and excluding these items in determination of a summary score means that the TESS does not have standard content across individuals. This raises the question of whether the patient-specific scores across individuals mean the same thing. The patient-specific index was initially used by MacKenzie (1986), with Tugwell (1987) adopting similar methodology in the MACTAR
and Law (1990) doing so in the Canadian Occupational Performance Measure. Unlike the proposed use of the TESS for both individuals and groups, MacKenzie (1986), Tugwell (1987) and Law (1990) restricted the use of their measures to within subject changes.

Whether we are truly measuring the "same" thing, that is physical disability, with the TESS when there is not a standard set of items across individuals is unanswerable at the present time. Theoretically, this question is answerable if it were known that two items represented a similar metric of difficulty for a given population. These items would then be interchangeable (McDowell, 1996). What is known, however, is the effect of the "not applicable" responses on the summary score of the TESS.

Summary scores for the TESS were calculated by excluding items rated as "not applicable" and alternately coding the "not applicable" responses with: "1-impossible to do", the mean difficulty, and "5-not at all difficult". There was no difference in the summary scores calculated by the different codings of "not applicable" (Davis, 1994). In order to maintain a patient-specific measure with face and content validity, the "not applicable" response has been maintained and the TESS summary score excludes "not applicable" items from the numerator and denominator.

Summary score calculation also requires decisions about item weighting. The weighting of items was evaluated by using equal item weighting, individual importance weights and group mean importance weights in determining the total score. There were no clinically meaningful differences in the total scores for the different algorithms tested as the difficulty and importance ratings are not statistically independent (Evans, 1991). Consequently, an equal item weight has been adopted and the total score is standardized out of one hundred with higher scores indicating lower levels of disability.

Psychometric techniques of internal consistency and factor analysis (Nunnally, 1978) were also used to evaluate how the items of the TESS related to each other (Davis, 1994; Davis, 1996). Cronbach's alpha (Cronbach, 1951) demonstrated that the lower extremity measure had high internal consistency (alpha=0.94). The theoretical construct of
the TESS is that it represents a single domain - physical function - and this was confirmed for the lower extremity measure by principal component analysis. Even after varimax rotation, all items loaded on the first factor accounting for 51.70% of the variance (Davis, 1994). Consequently, the psychometric results reaffirmed the theoretical concept that the TESS evaluates a single domain, physical function.

Test-retest reliability and preliminary validity were evaluated using a second sample of subjects. Test-retest reliability was evaluated pre-operatively and at three different times post-operatively with a forty-eight hour interval between tests. High levels of test-retest reliability were found for the lower extremity measure with intraclass correlation coefficients (Fleiss, 1986) of 0.92 to 0.97 (Davis, 1994; Davis, 1996).

Preliminary criterion validity was evaluated. The typical clinical picture for sarcoma is that of a patient presenting with relatively "normal" physical function. In the initial post-operative period function declines and then gradually improves over time. Means and standard deviations of the TESS scores from the lower extremity sample for these time periods produced the anticipated profile (Davis, 1994; Davis, 1996) suggesting that the measure is able to distinguish varying levels of disability.

Known group validity was evaluated by comparing the pre-operative TESS scores for bone versus soft tissue tumour subjects (Davis, 1994; Davis, 1996). Patients with bone tumours frequently present complaining of pain whereas patients with soft tissue lesions rarely present with symptoms of pain. As anticipated, bone tumour patients had statistically significant lower scores than patients with soft tissue tumours (lower extremity: mean=seventy-four and eighty-nine respectively, p<0.001, independent t-test).

The preliminary testing of the TESS suggests that its theoretical conceptualization, patient perceived physical function, is supported and that the TESS has adequate levels of reliability and validity. In particular, the TESS is feasible in the clinical setting as it is self-administered and takes only ten minutes to complete. The reliability coefficients meet the criteria for individual patient use. These features suggest that the TESS may be useful with
both groups and individuals (McHorney, 1995). However, the purposes of the TESS have been stated to include monitoring and evaluation of change and treatment effectiveness. The testing of the measure to date is insufficient to confirm the validity of the TESS.

Physical function is a hypothetical construct, the validity of which can only be evaluated by comparing the TESS scores to those of measures of related constructs and/or by confirmatory factor analysis. Consequently, the TESS needs to be further validated by replicating the previous factor analysis on a new sample, by evaluating construct validity and criterion validity. Determination of responsiveness of the TESS requires longitudinal study to determine the change detected over time and ultimately, response to treatment. Serial time-point evaluations will allow the calculation of effect sizes (Cohen, 1988) and comparison with other measures will allow calculation of the relative efficiency of the measure (Liang, 1990). Furthermore, the information obtained from the longitudinal evaluation of patients will have direct clinical relevance in that it will provide clinicians with information regarding the expected course of recovery over time. This will be valuable information for both clinicians and patients as it will allow informed education and expectations for patients undergoing limb preservation surgery for extremity sarcoma.

The current study has been designed to further develop and validate the TESS. The specific objectives are:

1) to confirm the item content of the TESS;
2) to evaluate the underlying content structure of the TESS using factor analysis;
3) to evaluate the construct validity of the TESS, including convergent and divergent validity; and,
4) to evaluate the criterion validity of the TESS, including known group validity and the ability of the TESS to detect change in status.

In summary, evaluation of existing generic cancer-specific and disease-specific health status measures demonstrate that there is not an adequate instrument according to
measurement theory criteria (Feinstein, 1987; Nunnally, 1978) to evaluate physical function for both groups and individuals of the extremity sarcoma population. Consequently, the TESS was conceptualized and has undergone preliminary development to provide a means of monitoring and evaluating change in physical function for this population. The content was determined based on clinician and patient perceptions of relevant items and the reliability and preliminary validity testing suggest that the TESS has potential to fulfill its intended purposes for groups and individuals. Further validity testing is required. The subsequent chapters describe the methods and results of the validity studies.
CHAPTER THREE
METHODS OF THE MEASUREMENT STUDY

The overall objective of this thesis was to develop and evaluate the measurement properties of the Toronto Extremity Salvage Score (TESS). Specifically, the objectives included confirmation of the item content and underlying content structure of the TESS and the evaluation of validity. The confirmation of content and validation of the TESS were evaluated in a longitudinal study in which a series of lower extremity sarcoma patients undergoing limb salvage surgery (with or without adjuvant irradiation or chemotherapy) were evaluated for functional status pre-operatively and at four times post-operatively. A small portion of the soft tissue sarcoma patients included in this longitudinal study were also participants in a randomized trial comparing the effectiveness of pre-operative versus post-operative irradiation in curable soft tissue sarcoma. As the secondary end points in this randomized trial included the same functional outcome measures completed at the same times as the measurement study for this thesis, it was possible to include these subjects in the longitudinal measurement study presented.

The longitudinal measurement study consists of two components: 1) the review of item content and scale structure, subsequently referred to as the Developmental Study; and, 2) the validation of the TESS, subsequently referred to as the Validation Study. The Developmental Study included evaluation of item generation and reduction, and factor analysis was used to evaluate the underlying item structure of the TESS. The Validity Study evaluated construct validity and criterion validity. As part of the validity evaluation
(Hays, 1992), responsiveness was evaluated by looking at the change in TESS scores over time following treatment. The study schema is presented in Figure 3.

**Study Objectives:** To confirm the content of the Toronto Extremity Salvage Score (TESS) and validate the measure.

**Routine Clinical Care**

**System Entry**

Patients enter specialized cancer care centre at Mount Sinai Hospital clinic, Princess Margaret Hospital clinic or London Regional Cancer Centre clinic on referral from outside centre physician. Staging and treatment planning finalized.

**0 weeks:** Treatment includes surgery with or without adjuvant irradiation or chemotherapy. All surgery is done at Mount Sinai Hospital.

**Cancer Treatment Follow-ups**

- **2 weeks:** routine post-operative care
- **6 weeks:** routine post-operative care
- **3 months:** routine post-operative/cancer care including chest x-ray.
- **6 months:** routine post-operative/cancer care including chest x-ray.
- **9 months:** routine post-operative/cancer care including chest x-ray.
- **12 months:** routine post-operative/cancer care including chest x-ray.
- **15 to 36 months:** routine post-operative/cancer care including chest x-ray.

**Study Process**

**Identification of Eligible Subjects**

Patient has non-metastatic disease and will be treated with limb salvage surgery.

**Study Entry:** Patient consents to enter study. Pre-operative functional status questionnaires completed.

**Evaluations for Functional Status**


**Data Analysis:** Data analysis for the Content and Validity Studies after the 100th subject has minimum one year follow-up post-surgery.

Figure 3 presents the schema of the longitudinal study in which the TESS content was verified and the measure was validated.
STUDY SETTING

Patients participating in this study were treated at the University Musculoskeletal Oncology Unit, Mount Sinai Hospital, Toronto and its affiliated clinics. Mount Sinai Hospital is a tertiary care centre and sarcoma patients from across Ontario are referred to the subspecialty Musculoskeletal Oncology Unit. A small number of patients are referred from outside the province, approximately 5% of the 300 new referrals seen per annum. About 100 new soft tissue sarcoma patients are seen per year with 65% of them presenting with primary extremity soft tissue sarcoma. Primary malignant and benign aggressive tumours of bone represent a further seventy new referrals per year with the remainder including patients with non-extremity lesions, metastatic bone tumours or benign lesions. (These data were obtained from the prospective database maintained at the University Musculoskeletal Oncology Unit, Mount Sinai Hospital, Toronto by permission of the Director, Dr. R.S. Bell). Patients under twelve years of age are referred to a pediatric centre.

Surgical management and some chemotherapy is provided at the Mount Sinai Hospital but adjuvant therapies and patient assessments may also be done at affiliated clinics. The affiliated clinics include the Sarcoma Clinic at the Princess Margaret Comprehensive Cancer Centre, Toronto (soft tissue sarcoma patients only) and the London Regional Cancer Centre (bone and soft tissue patients). Patients attend the London clinic based on residence and only patients with soft tissue sarcoma who require adjuvant radiotherapy attend the Princess Margaret Clinic. All others are followed at Mount Sinai Hospital. Following surgery, routine evaluation occurs at six weeks to evaluate tissue healing and patient recovery from surgery. Evaluation for local recurrence and systemic relapse is done at three monthly intervals from surgery for the next three years. The highest risk of recurrence is in the first three years (Bramwell, 1994; Harrison, 1993; Pisters, 1996).
PATIENT SAMPLE

Patients participating in the study were followed longitudinally with functional status measures (TESS, SF-36, MSTS (1987,1993) taken at baseline (prior to definitive surgery), six weeks, three, six and twelve months post-surgery. These times were chosen for repeated evaluation as clinical judgement suggests that they represent clinically significant benchmarks. At six weeks post-operatively, functional status is most influenced by the surgical intervention with soft tissue and bony healing in progress. Healing times are delayed by the use of adjuvant irradiation and chemotherapy compared to the times common in most orthopaedic procedures e.g. soft tissue healing in two to three weeks or fracture healing at six weeks versus two to three times longer for soft tissue and bone healing following irradiation or chemotherapy. Consequently, at three months post-operatively, many patients are only beginning progressive weight-bearing from crutches, requiring extended periods to protect extensive bone reconstructions. Patients on chemotherapy usually complete their final post-operative chemotherapy at about three months and, therefore, have further delays in bone healing and functional recovery. The six month point represents a time when patients, other than those delayed by complications, should be fully weight-bearing and reintegrating into their usual activities. Consequently, it is expected that the first six months following surgery will represent the greatest magnitude of recovery. However, in order to determine whether patients continue to improve, functional status measures were also taken at one year.

Patients were eligible for the longitudinal study if they met the following criteria:

1) primary malignant or benign aggressive tumour of the lower extremity;
2) age range 12 to 85 years;
3) the definitive treatment program had not been instituted at the time of registration in the study. The subject may have undergone biopsy or resection with positive margins elsewhere and presented for definitive surgery;
4) treatment included limb salvage surgery;
5) informed consent; or, for subjects under the age of 16 years, a guardian was asked to give consent; and
6) subjects with cognitive deficits were excluded from the study.

SAMPLE SIZE AND POWER CALCULATIONS

In choosing a sample size for the study, two factors were considered. The first was that the sample size was sufficient to detect a given amount of change in the TESS with minimum 80% power. The second consideration was based on the recognized heterogeneity of the sarcoma population. In recognition of this heterogeneity, the intent was to recruit 100 lower extremity subjects with bone and soft tissue tumors. Although the number is arbitrary and based on feasibility, it was felt that this sample was required to adequately represent the variability of site and reconstructive options for the lower extremity group. As the sample size was established, power calculations were estimated from previously collected data (Davis, 1994).

Effect sizes are typically calculated for independent samples in which the difference in the means of the two independent samples is divided by the pooled variance of the scores (Cohen, 1988). However, when using paired data or data from repeated measures the correlation between the measures needs to be considered (Cohen, 1988). This occurs by calculating the effect size for paired data by dividing the difference in the scores (change score) by the standard deviation of the difference. The effect of the correlation is such that the standard deviation of difference increases above the average of the standard deviations when $r$ is less than 0.50 and, alternately, the standard deviation of the difference decreases below the average of the standard deviation when $r$ is greater than 0.50. Cohen (1988) has shown that there is equivalency between the effect sizes of paired data and independent data adjusted by the correlation as shown below:

$$\text{independent effect size} = \frac{\text{change score}}{\text{standard deviation of difference}}$$
where the independent effect size is calculated by the difference of the two means divided by the pooled variance and $r$ equals the correlation of the measures at the two points in time (Cohen, 1988).

The power calculations for the current study were based on the previous knowledge that the smallest change in TESS scores for patients undergoing treatment for lower extremity sarcoma occurred between the six and twelve month post-operative times. Using Master’s thesis data (Davis, 1994) from patients at six months post-surgery, thirty-one subjects were followed to one year to provide pilot data for sample size and power calculations for the current study. The mean TESS score at six months was 73.6, $sd=16.17$ and at twelve months the mean was 79.2, $sd=15.65$. The TESS scores at six and twelve months were correlated, $r=0.33$. The paired effect size based on this pilot data was 0.43, of moderate magnitude according to Cohen (1988). A sample size that was able to detect this moderate change would be able to detect the expected larger changes between other time points for a given alpha and beta.

Based on the paired effect size, $alpha=0.05$, 100 subjects would provide 90% power to detect a moderate effect size of 0.43 (Cohen, 1988).

**ADMINISTRATION OF THE INSTRUMENTS**

Ethics approval for the study was obtained from the Human Subjects Review Committee, University of Toronto and, in accordance with the study protocol, patients were introduced to the study by a trained study assistant. If the patient agreed to participate in the study, they were given a package of questionnaires to complete. Initial contact occurred most frequently in one of the clinics. There were, however, a few occasions when patients were admitted directly to hospital resulting in the first contact occurring when the subject was an inpatient. The study assistant explained the purpose of the various questionnaires and provided instructions for completing the forms. The questionnaires were all self-administered and included demographics, the TESS, the SF-
36, and a global rating of disability. Questionnaire packages were collated such that the order of the instruments appeared to the subjects in random order. The clinical measures i.e. range of motion, strength, joint stability, and deformity necessary to score the MSTS (1987) were done by a single physiotherapist who was blinded to the results of the other measures. Follow-up measures were taken at six weeks, three, six and twelve months post-operatively as patients returned for their routine follow-up visits. These time-points were chosen as they coincided with routine follow-up but also because they coincided with critical times for bone and soft tissue healing.

INSTRUMENTS

The instruments used in the study included a socio-demographic profile, documentation of ambulatory aids, pain medication, the TESS, the SF-36, the Musculoskeletal Tumor Society Rating Scale (MSTS) 1987 and 1993 versions (Enneking, 1987; Enneking, 1993) and a global rating of disability.

As previously noted, the SF-36 is a generic health status measure and includes the subscales of physical function, role function-physical, bodily pain, general health, vitality, social function, role function-emotional, and mental health (Ware, 1992; Ware, 1993). A detailed description of the development of the SF-36 is presented in Appendix Two. Of particular interest for testing with the TESS were the domains of vitality, physical function, social function and mental health. Fatigue was a frequently reported symptom of the extremity sarcoma population (Davis, 1994; Davis, 1996). Therefore, the vitality domain of the SF-36, as it includes the concept of “fatigue” (Ware, 1993), was expected to be related to the TESS scores. Similarly, relationships between the physical, social and emotional function domains of the SF-36 and the TESS were evaluated.

Both the 1987 and 1993 versions of the MSTS (Enneking, 1987; Enneking, 1993) were evaluated in relation to the TESS. (These instruments were described in detail in Chapter Two. The 1987 version contains content heavily weighted to impairment
measures such as range of motion, strength, and joint stability (WHO, 1980) and is specific to an anatomical location e.g. pelvis/hip/proximal thigh. The rating criteria vary according to the anatomical location of the tumour. The 1993 version of the MSTS (Enneking, 1993) is evaluated based on the entire extremity as opposed to a regionalized anatomical location. The content of the 1993 MSTS is also weighted towards impairment measures but the items are more representative of functional limitations with such inclusions as walking and use of gait aids. Consequently, the two versions of the MSTS (Enneking, 1987; Enneking, 1993) provide different information.

The global ratings of disability are self-ratings by the subject and are based on a five point Likert-type scale. This global disability rating served as criterion measure against which the TESS scores were compared. The framing of the global ratings of disability was the same as the TESS items with the response options ranging from "not at all disabled" to "totally disabled".

ANALYSES

Data Quality

All questionnaires were manually reviewed for completeness by an individual not responsible for data collection. A random sample of five percent of each of the questionnaires was double data entered and compared for potential data entry errors. A maximum of two entry errors per questionnaire was set as the standard for proceeding to double data entry of the complete data set. Of the data reviewed, there were four errors noted: one gender coding error and one age error in the same subject; one MSTS (1987) error; and one error in an item of the mental health subscale of the SF-36 in two different subjects. These latter two errors was detected by out of range values on descriptive statistics. Consequently, data quality was such that complete double entry was not required.
Descriptive Statistics

Descriptive statistics consisting of means, standard deviations and frequencies were computed as appropriate for demographic and clinical descriptors for the sample. Descriptive statistics were also calculated for the TESS, SF-36 subscales, the two MSTS scales and the disability global rating scale.

All analyses for the thesis were conducted using the Statistical Package for the Social Sciences (SPSS version 6.1)\(^1\) or Stata Statistical Software (STATA 4.0)\(^2\).

**DEVELOPMENTAL STUDY**

**Item Generation and Reduction**

The intent of the content study was to confirm the inclusion of items as previously determined (Davis, 1994; Davis, 1996) in the lower extremity TESS questionnaire. Specifically, the intent was to evaluate, based on the responses to the open-ended questions, whether any additional items needed to be added to the questionnaire or alternatively, based on the importance ratings and "not applicable" responses, whether the number of items in the questionnaire could be reduced. A primary clinimetric strategy (Feinstein, 1987) was used for determining content of the questionnaires. This method used clinical judgement and patients' perceptions of relevant items in determining the content of the instrument, irrespective of how the items might relate to one another; the latter representing item generation and reduction based on a psychometric approach (Kerlinger, 1986; Nunnally, 1978). The methods and analyses used in the content study were the same as those from the initial developmental study (Davis, 1994; Davis, 1996).

Data from the six month questionnaires were analyzed for item content as previous work suggests that improvements in physical function plateau after six months (Davis, 1996). In evaluating the generation of new items, the open-ended questions were

---


\(^2\) StataCorp. 1995. Stata Statistical Software: Release 4.0 College Station, TX: Stata Corporation.
reviewed and items representing similar constructs were assigned to a single category e.g. the additional items were classified as sporting activities rather than including individual sports. Frequencies of the items in the open-ended questions were calculated and any item identified by thirty percent of the sample was added to the questionnaire. The *a priori* cut-point of thirty percent is the same as in the initial study (Davis, 1994; Davis, 1996).

Importance ratings and "not applicable" ratings were used to reduce items. An item that was rated as "totally unimportant" or "not applicable" by thirty percent of the respondents was removed from the questionnaire.

This methodology replicated that of the previous study (Davis, 1994; Davis, 1996) and was designed to verify the previous results. Consequently, it was anticipated that there would be no item reduction or generation from this clinimetric methodology.

**Internal Consistency and Factor Analysis**

**Coding of the TESS "not applicable" Responses**

The difficulty responses for each item of the TESS ranges from one ("impossible to do") to five ("not at all difficult"). There is a further option in that a subject may indicate that an item is "not applicable". The items coded "not applicable" are excluded from both the numerator and denominator in the calculation of the TESS summary score (actual score minus minimum possible score divided by the possible score range) for all analyses.

However, the evaluation of internal consistency and factor analysis requires that coding decisions are made about the "not applicable" response for individual items. The "not applicable" response can be coded as missing with the resultant loss of information as subjects are excluded from the internal consistency and factor analysis procedures as soon as an individual has a single missing response. (This option provides some confusion in separating "not applicable" responses from truly missing data, although this is not an issue for the current study as none of the completed TESS questionnaires had omitted
responses.) An alternate value may be used to represent the "not applicable" response designating it as separate from the one to five coded response options. This allows the entire sample to be maintained in the analysis.

Choice of either of the above options is debatable. The meaning of a "not applicable" response can be questioned as truly meaning that it is an activity that, by choice, is not a part of an individual's routine activities versus the response representing an activity the individual is unable to do. Coding the "not applicable" responses as "1-impossible to do" is representative of this latter scenario and would tend to over-estimate disability levels. As a corollary, coding the "not applicable" responses as "5-not at all difficult" represents the scenario of no disability for the activity and would underestimate the disability level.

Previous work (Davis, 1994; Davis, 1996) demonstrated that coding the "not applicable" response using "1-impossible to do", "5-not at all difficult" or using the mean difficulty rating for the item did not change the internal consistency or factor analysis results. However, to verify this finding, Cronbach's alpha and the factor analysis in the current study were repeated using the following codings for the "not applicable" response: zero; "1-impossible to do"; and "5-not at all difficult."

**Internal Consistency**

The second objective of the Content Study was to evaluate the underlying internal relationship of the items in the TESS to each other by psychometric methods. Psychometric evaluation of the item difficulty responses at six months post-operatively included evaluation of the item correlation matrix, internal consistency and factor analysis. The frequency of the response options for each item and the mean response were calculated as well as the item total correlations. Using the criteria of Ferketich (1991) in analyzing the item correlation matrix, the proportion of item-to-item correlations that
ranged between 0.30 and 0.70 and the average correlation of every item with every other item were determined. The mean item-to-item correlation should be 0.4 or higher.

The internal consistency of the questionnaires was evaluated by Cronbach’s alpha (Cronbach, 1951). Alpha provides a measure of the strength of the relationship amongst the items in the questionnaire. The stronger the relationship, i.e. the items are contributing to the same concept, the higher the alpha value. Alpha ranges from 0.0 to 1.0 and Nunnally (1978) recommends an alpha of 0.90 for measures used in clinical research. A more rigorous criteria of 0.95 is recommended if a questionnaire is to be used in making treatment decisions for individuals.

The TESS has been shown to have high internal consistency with alpha=0.94 (Davis, 1994; Davis, 1996) and it was anticipated that, due to the number of items in the measure, the current data would, at minimum, achieve this alpha.

**Factor Analysis**

Factor analysis is a multivariable statistical method that analyzes the underlying structure of the interrelationships among a number of variables to identify the fewest number of latent factors that explain the underlying meaning or hypothetical construct of the data (Kleinbaum, 1988; Nunnally, 1978; Nunnally, 1994; Zeller, 1980). The number of factors is smaller than the number of initial variables and is in keeping with the intent of providing a more parsimonious explanation of the data. Consequently, factor analysis serves two main purposes. One is to determine if a smaller number of variables can be used to explain the data. The second and primary purpose of its use in this thesis is to determine the underlying dimensions or constructs being measured by the data. The investigation of underlying dimensions of the variables included in the TESS can be considered part of the evaluation of the construct validity of the measure.

Principal Component Analysis (PCA) is a type of factor analysis that relies on the fewest assumptions. Its purpose is to explain as much of the total variance in the data as
possible with the minimal number of factors. The factors generated are a linear combination of an underlying hypothetical construct and the general form is: \( F = w_1 x_1 + w_2 x_2 + \ldots + w_i x_i \), where \( w \) = weight and \( x \) = variable. The weights are calculated such that each successive factor accounts for a smaller portion of the total common variance. The factors generated are uncorrelated. In PCA, as the total variance is accounted for, the total number of factors is equal to the number of variables and the total variance is equal to the number of variables.

The correlation matrix is most commonly used in factor analysis (Kleinbaum, 1988; Norman, 1993). In a typical correlation matrix, the diagonal has unities and the analysis using this unadjusted matrix therefore represents the relationships of the off-diagonal components of the matrix. Unities on the diagonal provide an analysis of the original variables in their entirety, including the common, unique and error components of the variance. In using PCA and evaluating the variable relationship based on total variance, unities will be maintained in the correlation matrix.

The ultimate goal of factor analysis is to obtain simple structure (Kleinbaum, 1988), that is, to obtain results in which each variable relates highly to only one factor and to obtain a smaller number of factors (representing dimensions of the construct) than the original number of variables. The first analysis solution provides a general solution that may produce many, still difficult to interpret factors. If PCA has been used, the number of factors generated will be equal to the number of variables. By convention (Nunnally, 1978; Nunnally, 1994; Norman, 1993), factors are maintained if they have an eigenvalue of one. An eigenvalue is the proportion of variance explained by the factor divided by the number of variables. Consequently a factor with an eigenvalue of less than one explains less than the variance of one variable. Variables are assigned to a factor on the basis of highest factor loadings. Factor loadings are the correlation of the variable to the factor and Nunnally (1978; 1994) suggests that a minimal loading of 0.35 is required to assign a variable to a factor. This initial analysis may result in variables that have "high" loadings
on more than one factor such that the results are difficult to interpret. In an attempt to obtain simple structure as defined above, the factors retained from the unrotated analysis are re-analyzed using rotation.

Orthogonal rotation rotates the factors such that they remain uncorrelated. Orthogonal rotation has the advantage, by maintaining uncorrelated factors, that the variable correlations with the factor (factor loadings) are the same as the weights in the linear equation for the factors. The variance accounted for by each factor remains the same even after orthogonal rotation. The purpose of rotation is to obtain structural simplicity, the criteria of which are variables related highly to only one factor and factors defined by a small number of variables. It is important to recognize that structural simplicity does not guarantee conceptual clarity or meaningfulness. Consequently, interpretation of the results must be undertaken on the basis of interpretable and meaningful factors (Zeller, 1980). This is particularly important in interpreting the results of PCA for the TESS as this analysis is part of the validation of the instrument.

The difficulty ratings from the six month scores of the TESS were used in the PCA as this time represented the minimum criteria for subject inclusion in the sample upon which factor analysis was done (Davis, 1994). The unadjusted correlation matrix was used in the analysis and factors were retained on the basis of the eigenvalue equal to one criteria (Nunnally, 1978; Nunnally, 1994; Norman, 1993). Items were considered to load on a factor if the factor loading was 0.35 or higher (Nunnally, 1978; Nunnally, 1994) and varimax orthogonal rotation and oblique rotation were used in an attempt to obtain simple structure.

Previous factor analysis suggested that the TESS represented a single factor - physical function - that accounted for 52% of the variance in the data (Davis, 1994; Davis, 1996). This single factor resulted even after varimax rotation of the data. However, six of the items of the TESS loaded on more than one factor such that true simple structure was not attained. Similar results were anticipated with the current analyses in the hope that
there would be fewer items loading on multiple factors as the sample size was increased by one third.

**VALIDITY STUDY**

**Construct Validity**

The TESS is conceptualized as a disability measure of physical function and should, therefore, have predictable positive relationships with other measures of physical function and related constructs, such as social and mental health (Aaronson, 1988; Aaronson, 1991; Bergner, 1976; Jette, 1986b; Meenan, 1982). Verification of these relationships would support the construct validity of the TESS.

In particular, the TESS and the physical function subscale of the SF-36 should have a strong positive correlation (r>0.80) as they have many items in common. The MSTS 1987 and 1993 versions have also been cited as measures of physical function. However, the 1987 version of the MSTS contains many impairment items (e.g. pain, range of motion, strength, joint stability and joint deformity), representing five of the seven items. The TESS is a disability measure and the relationship of these impairment and disability measures is expected to be positive but only in the moderate range of 0.40 to 0.80 (Badley, 1987; Meenan, 1982; Spiegel, 1988). The 1993 version of the MSTS shares three items with the 1987 version, pain, overall function, and emotional acceptance of the surgery. The remaining items include use of ambulatory aids, walking ability and a dimension named "gait handicap". The 1993 MSTS has less impairment focus but still does not represent the breadth of disability of the TESS. Therefore, the MSTS 1993 version should be positively correlated with the TESS and it is reasonable that the correlation coefficients, while moderate in magnitude, would be higher than those for the MSTS 1987 version and the TESS.

Social and mental health are different but related constructs when evaluated in relation to physical function. This relationship is demonstrated by moderately positive
correlations between social and mental health and physical function. The correlations of the TESS with the SF-36 subscales of social function and mental health are expected to support the relationship of these constructs, that is social and mental health with physical function.

Spearman's correlation coefficient was used to evaluate the relationship of the TESS with the SF-36 subscales and the MSTS 1993. The TESS/MSTS 1987 relationship was evaluated using Kendall's tau (Kendall, 1990). Coefficients were defined as low if less than 0.40, moderate if in the range of 0.40 to 0.80 and high if greater than 0.80 (Nunnally, 1978). Correlations were calculated for the scores pre-operatively and at six weeks, three, six and twelve months post-operatively.

**Criterion Validity**

The global rating of disability was also evaluated against the lower extremity TESS to determine if the TESS scores discriminated amongst subjects' self-ratings of overall disability. The data were analyzed using a one-way analysis of variance. The means with 95% confidence intervals for each time point were calculated.

The post-operative use of walking aids was used as a criterion measure against which to evaluate the scores of the lower extremity TESS. The use of ambulatory aids six weeks post-operatively is largely influenced by the surgeon's preference for patient weight-bearing and depends on the extent of reconstruction and the use of chemotherapy. For example, patients treated by curettage and bone graft may have protected weight-bearing for three months despite overall improvements in their general function. Due to the clinician imposed use of ambulatory aids, it was anticipated that six weeks post-operatively when recovery from surgery is the major influence, and at twelve months post-operatively when TESS scores are anticipated to have plateaued would be the time points at which TESS scores would most likely be discriminated by use of ambulatory aids. The three and six month post-operative TESS scores should be less successful at
discriminating use of ambulatory aids. These hypotheses were evaluated by presenting mean TESS scores for the ambulatory aids at each post-operative time.

Clinical experience suggests that lower extremity sarcoma patients present with relatively normal physical function pre-operatively, although bone patients have been shown to have lower scores pre-operatively than patients with soft tissue sarcoma (Davis, 1994; Davis, 1996). Function declines at six weeks post-operatively and then gradually improves over the subsequent months. The mean scores for the TESS, SF-36 physical function domain and the two versions of the MSTS should reflect this pattern and the scores are presented for each of the measures at the five time points. An independent t-test was used to evaluate the pre-operative TESS scores for the bone versus soft tissue patients to confirm the previous finding (Davis, 1994; Davis, 1996).

The TESS has been developed to detect changes in physical function over time in response to treatment (monitoring purpose), surgery with or without adjuvant therapy, and as an outcome measure in clinical trials. Validation of the TESS, therefore, requires that its responsiveness be demonstrated (Hay, 1992).

The majority of extremity sarcoma patients present pre-operatively with relatively normal scores, the exception being the less than 1% (Mount Sinai Hospital, Sarcoma Data Base, 1989-1995) of patients who present with pathological fractures. (Patients presenting with pathological fractures will have very low pre-operative functional scores.) Pathological fracture in extremity sarcoma is not a contra-indication to limb salvage surgery and there are no data to suggest that patients presenting with primary sarcoma who undergo limb salvage surgery have inferior functional status when compared to others. Six weeks, three, six and twelve months post-operatively were chosen for times of follow-up as they represent significant clinical milestones. At six weeks post-surgery, the TESS scores are largely influenced by the magnitude of the treatment in conjunction with post-surgical recovery. At subsequent times, the impact of recovery from having surgery is lessened and the TESS scores reflect limitations in physical function. This anticipated
pattern of recovery was evaluated graphically and by evaluating descriptive statistics of the TESS over time for the total sample and for subgroups of subjects.

Mean TESS scores over time for proximal and distal thigh soft tissue tumours, bone tumour patients treated by curettage and bone graft and bone tumour patients treated with endoprostheses at the knee were evaluated. These sub-groups were chosen on the basis of feasibility in representing anatomical site and reconstruction. These data sets have the potential to characterize physical function of specific patient groups at designated times such that patient education might be improved. Potential outliers might also be identified such that timely intervention may be undertaken. Means and standard deviations were calculated over time. Future validation of these developmental data could be conducted using a new group of patients.

Differences in outcome between cases and controls in clinical trials are evaluated using effect sizes (Cohen, 1988). The standardized effect size (difference in TESS score divided by the standard deviation of the change) was calculated between six weeks and three months, three and six months, six and twelve months and the pre-operative and twelve month scores (Cohen, 1988).

As a final step in evaluating the ability of the TESS to detect change, the relative efficiency of the TESS compared to the MSTS (1987, 1993 versions) and the physical function subscale of the SF-36 were evaluated. Relative efficiency was calculated by the ratio of the effect sizes of the TESS to the comparative measure (Liang, 1985). An effect size of greater than one would indicate that the TESS is more efficient (i.e. requires a smaller sample size for a given power and a given effect size) than the comparative measure (Liang, 1985).
CHAPTER FOUR

RESULTS

INTRODUCTION

This chapter presents the results of the development (Developmental Study) and tests the validity (Validity Study) of the Toronto Extremity Salvage Score (TESS). In particular, item reduction, item generation, internal consistency and factor analysis results are presented within the context of the Developmental Study. The Validity Study reports the results of construct validity, criterion validity and responsiveness.

PATIENT SAMPLE

One hundred and six eligible lower extremity subjects consented to participate in the study. Nine subjects were permanently lost as outlined in Tables 1a and 1b, leaving a total of ninety-seven subjects included in the study. The permanent losses were the result of a subject who died of a myocardial infarction four months following surgery, six subjects who developed metastatic disease within the first year following surgery and two subjects who were missing evaluations at more than one time point as they are followed out of province.

As is common in clinical research requiring evaluations at multiple time points, there were missing data. Eight individual subjects missed a single evaluation (less than two percent of the total evaluations) as shown in Tables 1a and 1b. The repeated measures
are required predominantly for the responsiveness analysis and imputation was undertaken with the intent of minimizing bias for this analysis. The values imputed for the summary score of each of the measures was the average of the summary score of the evaluation preceding and following the missing time point. Only the summary score for the measure was imputed; individual item values were not imputed for any of the measures.

TABLE 1

LOWER EXTREMITY: ELIGIBLE SUBJECTS LOST TO FOLLOW-UP

1a: Subjects Excluded from Analysis

<table>
<thead>
<tr>
<th>Reason for loss</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-disease/treatment related death</td>
<td>1</td>
</tr>
<tr>
<td>Development of metastatic disease within year of follow-up:</td>
<td></td>
</tr>
<tr>
<td>currently alive with disease</td>
<td>4</td>
</tr>
<tr>
<td>deceased of disease</td>
<td>2</td>
</tr>
<tr>
<td>Eligible loss as follow-up out of province</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9</strong></td>
</tr>
</tbody>
</table>

1b: Number of Subjects with Missing Evaluations Per Time Point
(Total of 8 patients, each at a single time point)

<table>
<thead>
<tr>
<th>Time point</th>
<th>Pre-operative</th>
<th>6 weeks</th>
<th>12 weeks</th>
<th>6 months</th>
<th>12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: one subject has been retained in the study with missing MSTS evaluations at 2 time points. Subject was followed at Northern Ontario Cancer Centre and only patient-completed questionnaires were completed.
The ninety-seven lower extremity subjects included in this study (Table 2) consisted of fifty-four males and forty-three females with a mean age of forty-three years. This gender ratio and age are typical of a sarcoma population as is the anatomical distribution of the tumours with the proximal extremity more common than the distal.

TABLE 2

LOWER EXTREMITY SAMPLE CHARACTERISTICS
(n = 97)

<table>
<thead>
<tr>
<th>Age</th>
<th>mean = 42.7, sd=19.8, range 13-82</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Region</td>
<td>proximal to knee joint</td>
</tr>
<tr>
<td></td>
<td>distal to knee joint</td>
</tr>
<tr>
<td>Site</td>
<td>pelvis/hip/proximal thigh</td>
</tr>
<tr>
<td></td>
<td>distal thigh/knee/proximal leg</td>
</tr>
<tr>
<td></td>
<td>distal leg/ankle/foot</td>
</tr>
<tr>
<td>Tissue type</td>
<td>bone</td>
</tr>
<tr>
<td></td>
<td>soft</td>
</tr>
<tr>
<td>Tumour type</td>
<td>benign aggressive</td>
</tr>
<tr>
<td></td>
<td>malignant</td>
</tr>
<tr>
<td>Treatment</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>surgery, irradiation</td>
</tr>
<tr>
<td></td>
<td>surgery, chemotherapy</td>
</tr>
<tr>
<td></td>
<td>surgery, irradiation, chemotherapy</td>
</tr>
<tr>
<td>Reconstruction excision</td>
<td></td>
</tr>
<tr>
<td></td>
<td>excision, tissue transfer</td>
</tr>
<tr>
<td></td>
<td>curettage, bone graft</td>
</tr>
<tr>
<td></td>
<td>endoprostheses</td>
</tr>
<tr>
<td></td>
<td>other</td>
</tr>
<tr>
<td></td>
<td>54.7%</td>
</tr>
<tr>
<td></td>
<td>64.9%</td>
</tr>
<tr>
<td></td>
<td>35.1%</td>
</tr>
<tr>
<td></td>
<td>45.3%</td>
</tr>
<tr>
<td></td>
<td>40.2%</td>
</tr>
<tr>
<td></td>
<td>14.4%</td>
</tr>
<tr>
<td></td>
<td>45.4%</td>
</tr>
<tr>
<td></td>
<td>54.6%</td>
</tr>
<tr>
<td></td>
<td>38.1%</td>
</tr>
<tr>
<td></td>
<td>61.9%</td>
</tr>
<tr>
<td></td>
<td>42.3%</td>
</tr>
<tr>
<td></td>
<td>46.3%</td>
</tr>
<tr>
<td></td>
<td>10.3%</td>
</tr>
<tr>
<td></td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td>59.8%</td>
</tr>
<tr>
<td></td>
<td>8.2%</td>
</tr>
<tr>
<td></td>
<td>17.5%</td>
</tr>
<tr>
<td></td>
<td>10.3%</td>
</tr>
<tr>
<td></td>
<td>4.1%</td>
</tr>
</tbody>
</table>

*Twenty-six of these subjects are participants in a randomized trial sponsored by the National Cancer Institute of Canada-Clinical Trials Group

region. Approximately two-thirds of the subjects had tumours that were malignant (with metastatic potential) and one-third had benign aggressive lesions that, while not life
threatening, have a high risk of local recurrence. Over half the sample was treated with surgery and adjuvant therapy (either irradiation or chemotherapy) and the surgical procedures represent a variety of reconstructive procedures. The high number of patients treated with surgery and irradiation and classified as excision in the reconstructive category reflects the number of soft tissue tumour subjects included in the sample. Subjects with bone tumours are included in smaller numbers in the excision group and include all the subjects who were reconstructed with metallic prostheses as well as patients who were curetted and bone grafted. Twenty-six of the subjects with soft tissue tumours are also participants in a multi-centered randomized clinical trial sponsored by the National Cancer Institute of Canada-Clinical Trials Group.

Clinical experience suggests that, of patients experiencing no wound or reconstructive complications, those treated by chemotherapy and endoprosthesis experience the most physical disability. Patients treated for a large, soft tissue sarcoma of the thigh who lose a significant portion of functioning quadriceps, while having residual physical deficit will be less affected than those with endoprostheses. Patients treated by curettage and bone graft while taking three months to progress their weight-bearing have greater potential to return to full activity. Consequently, while patients are treated for the same disease - sarcoma - recovery of physical function may vary according to the treatment modalities and reconstructive procedures used.

Co-morbid diseases may also affect physical function. Eleven subjects each reported a single co-morbid disease. Blinded to the functional scores of the subjects, the two surgeons who treated this patient population determined that four individuals had co-morbid disease that would impact upon their function. These included a seventy-seven year old woman who is a legally blind, insulin-dependent diabetic; a seventy-two year old woman who is on oxygen for chronic obstructive pulmonary disease; a sixty-six year old man with severe osteoarthritis of his knees who underwent total knee arthroplasty within six months of his tumour surgery; and a seventy-one year old gentleman who required
open heart surgery prior to his tumour surgery. The low prevalence of co-morbid disease is as expected considering the relatively young age of the majority of the sample (mean of 42.7 years). It is noteworthy, however, that the four individuals with significant co-morbid disease were older than sixty-years.

**DEVELOPMENTAL STUDY**

**Item Generation and Reduction**

The open-ended questions from the six months post-operative questionnaire were used in item generation. Fifteen subjects identified items most of which consisted of individual preference sport or leisure activities e.g. soccer, cards, carpentry. Of the fourteen items identified in these categories, only three items were identified more than once with two individuals listing the same items. Four further individual items: squatting, using public transit, lifting heavy objects, and laying on stomach were identified. None of the items identified met the thirty percent criteria for addition to the questionnaire.

Probing was not used in this study to elicit further items and the effect of probing for additional items is unknown. However, the initial pilot study included patient interviews in which the interviewer attempted to elicit further items (Davis, 1994) and the previous content work (Davis, 1994; Davis, 1996) also provided the opportunity for the addition of items. This prior work combined with the very low frequency of items generated in the current study, suggest that the content of the TESS is comprehensive of physical function and represents the perceptions of patients.

Item reduction was undertaken using the importance and "not applicable" responses of the six months post-operative data (n=94 as three subjects were missing evaluations for this time point). There were no missing item responses within the ninety-four questionnaires. Six items had high "not applicable" response frequencies although none of these items met the thirty percent criteria for exclusion from the questionnaire at six months post-operatively. The six items with the highest "not applicable" responses
were: sexual activity (28/94 or 29.8%), garden/yard work (18/94 or 19.1%), driving (18/94 or 19.1%), work duties (16/94 or 17.0%), work hours (17/94 or 18.1%), and sports (16/94 or 17.0%). The frequencies for the "totally unimportant" response were low with gardening/yard work having the highest response, with nine of ninety-four (9.6%) giving this response. The difficulty and importance response frequencies for each item are listed in Table 3.

The frequency of the "not applicable" responses at six weeks and one year were compared to the six months post-operative results to evaluate the consistency of the use of the "not applicable" response. Across times, the same items (sexual activity, garden/yard work, driving, work duties, work hours and sports) have the highest "not applicable" response frequencies. With the exception of work duties and work hours, a consistent number of individuals chose the "not applicable" response option at each time. Work duties varied from 16 to 23 "not applicable" responses and work hours varied from 11 to 25 "not applicable" responses depending on the time of evaluation. The high frequency of the "not applicable" item at six weeks post-operatively may be influenced by the number of individuals on sick benefits during the period such that completion of work duties over regularly scheduled hours is truly "not applicable". Overall, the consistency of the use of the "not applicable" response over time supports the findings of the six months time point.

The concern in having a "not applicable" response option is that the use of the response will become idiosyncratic rather than truly representing an activity that is not a usual activity in an individual's daily life. However, there is consistency in the frequency of use of the "not applicable" response by subjects across time. Furthermore, the items with the highest "not applicable" responses in the current data set are the same as those in the previous study (Davis, 1994) and this further suggests that selection of this response option is not idiosyncratic.
<table>
<thead>
<tr>
<th>Item</th>
<th>impossible to do</th>
<th>extremely difficult</th>
<th>moderately difficult</th>
<th>a little bit difficult</th>
<th>not at all difficult</th>
<th>not applicable</th>
<th>totally unimportant</th>
<th>mean difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>pants</td>
<td>0</td>
<td>0</td>
<td>9.6</td>
<td>24.5</td>
<td>66.0</td>
<td>0</td>
<td>3.2</td>
<td>4.6</td>
</tr>
<tr>
<td>shoes</td>
<td>2.1</td>
<td>1.1</td>
<td>11.7</td>
<td>30.9</td>
<td>54.3</td>
<td>0</td>
<td>1.1</td>
<td>4.4</td>
</tr>
<tr>
<td>socks</td>
<td>2.1</td>
<td>2.1</td>
<td>8.5</td>
<td>33.0</td>
<td>54.3</td>
<td>0</td>
<td>3.2</td>
<td>4.3</td>
</tr>
<tr>
<td>shower</td>
<td>3.2</td>
<td>1.1</td>
<td>8.5</td>
<td>11.7</td>
<td>73.4</td>
<td>2.1</td>
<td>3.2</td>
<td>4.5</td>
</tr>
<tr>
<td>light household</td>
<td>0</td>
<td>3.2</td>
<td>11.7</td>
<td>14.9</td>
<td>63.8</td>
<td>6.4</td>
<td>7.4</td>
<td>4.5</td>
</tr>
<tr>
<td>garden/yard</td>
<td>10.6</td>
<td>11.7</td>
<td>5.3</td>
<td>14.9</td>
<td>38.3</td>
<td>19.1</td>
<td>9.6</td>
<td>3.7</td>
</tr>
<tr>
<td>meal preparation</td>
<td>0</td>
<td>2.1</td>
<td>7.4</td>
<td>17.0</td>
<td>66.0</td>
<td>7.4</td>
<td>4.2</td>
<td>4.6</td>
</tr>
<tr>
<td>shopping</td>
<td>3.2</td>
<td>2.1</td>
<td>14.9</td>
<td>23.4</td>
<td>55.3</td>
<td>1.1</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>heavy household</td>
<td>9.6</td>
<td>8.5</td>
<td>10.6</td>
<td>16.0</td>
<td>50.0</td>
<td>5.3</td>
<td>8.5</td>
<td>3.9</td>
</tr>
<tr>
<td>bathtub</td>
<td>2.1</td>
<td>2.1</td>
<td>16.0</td>
<td>21.3</td>
<td>56.4</td>
<td>2.1</td>
<td>3.2</td>
<td>4.2</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>0</td>
<td>0</td>
<td>3.2</td>
<td>19.1</td>
<td>77.7</td>
<td>0</td>
<td>5.3</td>
<td>4.7</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>0</td>
<td>0</td>
<td>4.3</td>
<td>29.8</td>
<td>66.0</td>
<td>0</td>
<td>6.4</td>
<td>4.6</td>
</tr>
<tr>
<td>kneeling</td>
<td>18.1</td>
<td>12.8</td>
<td>12.8</td>
<td>24.4</td>
<td>30.9</td>
<td>1.1</td>
<td>5.3</td>
<td>3.4</td>
</tr>
<tr>
<td>bending</td>
<td>1.1</td>
<td>4.3</td>
<td>17.0</td>
<td>24.5</td>
<td>53.8</td>
<td>2.1</td>
<td>3.2</td>
<td>4.2</td>
</tr>
<tr>
<td>upstairs</td>
<td>0</td>
<td>9.6</td>
<td>17.0</td>
<td>22.3</td>
<td>51.1</td>
<td>0</td>
<td>1.1</td>
<td>4.1</td>
</tr>
<tr>
<td>downstairs</td>
<td>1.1</td>
<td>3.2</td>
<td>20.2</td>
<td>18.1</td>
<td>56.4</td>
<td>0</td>
<td>2.1</td>
<td>4.2</td>
</tr>
<tr>
<td>driving</td>
<td>4.3</td>
<td>1.1</td>
<td>5.3</td>
<td>10.6</td>
<td>59.6</td>
<td>19.1</td>
<td>3.2</td>
<td>4.4</td>
</tr>
<tr>
<td>walking indoors</td>
<td>1.1</td>
<td>0</td>
<td>4.3</td>
<td>16.0</td>
<td>78.7</td>
<td>0</td>
<td>3.2</td>
<td>4.7</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>3.2</td>
<td>3.2</td>
<td>13.8</td>
<td>21.3</td>
<td>58.5</td>
<td>0</td>
<td>3.2</td>
<td>4.3</td>
</tr>
<tr>
<td>sitting</td>
<td>0</td>
<td>3.2</td>
<td>1.1</td>
<td>12.8</td>
<td>81.9</td>
<td>1.1</td>
<td>2.1</td>
<td>4.7</td>
</tr>
<tr>
<td>hills</td>
<td>1.1</td>
<td>6.4</td>
<td>19.1</td>
<td>27.7</td>
<td>43.6</td>
<td>2.1</td>
<td>1.1</td>
<td>4.0</td>
</tr>
<tr>
<td>standing</td>
<td>0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>89.4</td>
<td>1.1</td>
<td>3.2</td>
<td>4.8</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>12.8</td>
<td>10.6</td>
<td>9.6</td>
<td>30.1</td>
<td>30.1</td>
<td>5.3</td>
<td>1.1</td>
<td>3.5</td>
</tr>
<tr>
<td>getting in/out of car</td>
<td>0</td>
<td>2.1</td>
<td>9.6</td>
<td>30.9</td>
<td>56.4</td>
<td>1.1</td>
<td>1.1</td>
<td>4.4</td>
</tr>
<tr>
<td>sex</td>
<td>3.2</td>
<td>3.2</td>
<td>4.3</td>
<td>12.8</td>
<td>46.8</td>
<td>29.8</td>
<td>5.3</td>
<td>4.3</td>
</tr>
<tr>
<td>work duties</td>
<td>4.3</td>
<td>3.2</td>
<td>10.6</td>
<td>20.2</td>
<td>44.7</td>
<td>17.0</td>
<td>5.3</td>
<td>4.2</td>
</tr>
<tr>
<td>work hours</td>
<td>10.6</td>
<td>1.1</td>
<td>9.6</td>
<td>17.0</td>
<td>43.6</td>
<td>18.1</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td>leisure activities</td>
<td>7.4</td>
<td>5.3</td>
<td>13.8</td>
<td>26.6</td>
<td>46.8</td>
<td>0</td>
<td>2.1</td>
<td>4.0</td>
</tr>
<tr>
<td>social activities</td>
<td>0</td>
<td>0</td>
<td>5.3</td>
<td>8.5</td>
<td>86.2</td>
<td>0</td>
<td>1.1</td>
<td>4.8</td>
</tr>
<tr>
<td>sports</td>
<td>17.0</td>
<td>10.6</td>
<td>14.9</td>
<td>14.9</td>
<td>25.5</td>
<td>17.0</td>
<td>4.3</td>
<td>2.7</td>
</tr>
</tbody>
</table>
In summary, the item generation and reduction results confirm the content of the TESS as anticipated and suggest that no further items should be added to the lower extremity TESS nor should any of the items be excluded.

**Internal Consistency and Factor Analysis**

**Coding of the TESS "not applicable" Responses**

Data coded as missing for internal consistency analysis and factor analysis results in those cases being dropped from the data set. There were no missing items (i.e. no subject failed to complete an item within a TESS questionnaire) at any of the evaluation times. However, as noted previously, there were two to three subjects who missed complete evaluations (i.e. all questionnaires) at six weeks, three months and six months post-operatively (Table 1b). It was unreasonable to impute thirty difficulty and thirty importance ratings for the TESS for these individuals and they were excluded in the internal consistency analysis and factor analysis.

As noted previously, the calculation of the summary score for an individual subject omits any items coded as "not applicable". However, fifty subjects would have been lost from the internal consistency and factor analysis if the "not applicable" response was coded as missing, representing a significant loss of information. Consequently, the analyses were repeated using three different codings for the "not applicable" response: 1) zero; 2) 1-impossible to do; and, 3) 5-not at all difficult. The coding of the "not applicable response" did not change the results of the mean item-to-item correlation within a given time point. The mean correlation of the matrix at six months post-operatively was: 0.41 with sd=0.15 with the "not applicable" response coded as zero; 0.47 with sd=0.16 with the "not applicable" response coded as one; and 0.51 with sd=0.15 with the "not applicable" response coded as five. In fact, the mean correlation of the matrix was consistent over multiple time points for each of the three codings of the "not applicable" response (range 0.41 to 0.52 with standard deviations ranging from 0.15 to 0.17).
Similarly, the Cronbach's alpha was consistent (alpha ranging from 0.94 to 0.96) as were the factor analysis results (i.e. simple structure was not achieved). Due to the similarity of the results, only the six month post-operatively internal consistency and factor analysis results are presented in detail in the body of the thesis.

The coding of the "not applicable" response did not impact on the results for two reasons. The first is that review of the item-to-item correlations shows that the values are not only consistent within and between times with three different coding values but that the mean item-to-item correlation is around 0.50 for most items. The exceptions are sexual activity and driving. Sexual activity had only seven of thirty items with item-to-item correlation in the range of 0.30 to 0.70 and driving had only four of thirty items in this range. Consequently, the mean item-to-item correlations for these two items was low, 0.30 and 0.29 for sexual activity and driving respectively. This consistent mid-range value neither increases nor decreases the value of alpha nor does it change the factor analysis.

The second reason that the "not applicable" coding did not have an impact is because few individuals chose this response in a small number of items in the thirty item questionnaire. Half the sample chose "not applicable" as a response on fewer than seven items. One subject chose eleven items as "not applicable". Forty-seven subjects did not have any "not applicable" responses. In summing the items and standardizing the score out of one hundred, each item makes a small contribution to the total score. With the low frequency of the "not applicable" items per individual, the coding of the response does not change the results.

**Internal Consistency**

Internal consistency evaluates how items relate to each other and to the total score and is based on item-to-item correlations. Table 4 presents the correlation matrix for the item responses of the six months post-operative lower extremity questionnaire and a summary of the correlation matrix is presented in Table 5. The item-to-item correlation for
### TABLE 4

**CORRELATION MATRIX SIX MONTHS POST-OPERATIVELY**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-pants</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-shoes</td>
<td>.73</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-socks</td>
<td>.60</td>
<td>.67</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-shower</td>
<td>.46</td>
<td>.48</td>
<td>.34</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-light household</td>
<td>.33</td>
<td>.29</td>
<td>.39</td>
<td>.34</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-garden/yard</td>
<td>.20</td>
<td>.37</td>
<td>.35</td>
<td>.37</td>
<td>.42</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-meal preparation</td>
<td>.30</td>
<td>.32</td>
<td>.32</td>
<td>.41</td>
<td>.64</td>
<td>.26</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-shopping</td>
<td>.50</td>
<td>.51</td>
<td>.46</td>
<td>.46</td>
<td>.45</td>
<td>.44</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-heavy household</td>
<td>.48</td>
<td>.49</td>
<td>.55</td>
<td>.63</td>
<td>.54</td>
<td>.44</td>
<td>.69</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J-bathtub</td>
<td>.67</td>
<td>.61</td>
<td>.55</td>
<td>.54</td>
<td>.42</td>
<td>.44</td>
<td>.48</td>
<td>.62</td>
<td>.56</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-in/out of bed</td>
<td>.39</td>
<td>.46</td>
<td>.37</td>
<td>.47</td>
<td>.37</td>
<td>.30</td>
<td>.35</td>
<td>.51</td>
<td>.47</td>
<td>.54</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-rising from a chair</td>
<td>.58</td>
<td>.61</td>
<td>.47</td>
<td>.47</td>
<td>.31</td>
<td>.18</td>
<td>.34</td>
<td>.18</td>
<td>.45</td>
<td>.50</td>
<td>.68</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-kneeling</td>
<td>.45</td>
<td>.43</td>
<td>.49</td>
<td>.46</td>
<td>.45</td>
<td>.47</td>
<td>.41</td>
<td>.61</td>
<td>.66</td>
<td>.50</td>
<td>.43</td>
<td>.43</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-bending</td>
<td>.41</td>
<td>.37</td>
<td>.36</td>
<td>.52</td>
<td>.43</td>
<td>.43</td>
<td>.43</td>
<td>.42</td>
<td>.62</td>
<td>.47</td>
<td>.41</td>
<td>.47</td>
<td>.57</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>O-upstairs</td>
<td>.55</td>
<td>.58</td>
<td>.58</td>
<td>.52</td>
<td>.48</td>
<td>.33</td>
<td>.50</td>
<td>.70</td>
<td>.68</td>
<td>.60</td>
<td>.50</td>
<td>.63</td>
<td>.58</td>
<td>.55</td>
<td>1</td>
</tr>
<tr>
<td>P-downstairs</td>
<td>.45</td>
<td>.48</td>
<td>.52</td>
<td>.43</td>
<td>.42</td>
<td>.38</td>
<td>.40</td>
<td>.75</td>
<td>.59</td>
<td>.55</td>
<td>.48</td>
<td>.58</td>
<td>.67</td>
<td>.48</td>
<td>.81</td>
</tr>
<tr>
<td>Q-driving</td>
<td>.04</td>
<td>.08</td>
<td>.08</td>
<td>.21</td>
<td>.01</td>
<td>.27</td>
<td>.05</td>
<td>.30</td>
<td>.22</td>
<td>.02</td>
<td>.28</td>
<td>.11</td>
<td>.23</td>
<td>.28</td>
<td>.14</td>
</tr>
<tr>
<td>R-walking indoors</td>
<td>.40</td>
<td>.33</td>
<td>.37</td>
<td>.60</td>
<td>.34</td>
<td>.26</td>
<td>.37</td>
<td>.53</td>
<td>.46</td>
<td>.47</td>
<td>.52</td>
<td>.43</td>
<td>.42</td>
<td>.45</td>
<td>.56</td>
</tr>
<tr>
<td>S-walking outdoors</td>
<td>.50</td>
<td>.51</td>
<td>.58</td>
<td>.60</td>
<td>.43</td>
<td>.35</td>
<td>.47</td>
<td>.76</td>
<td>.63</td>
<td>.61</td>
<td>.62</td>
<td>.50</td>
<td>.58</td>
<td>.45</td>
<td>.74</td>
</tr>
<tr>
<td>T-sitting</td>
<td>.30</td>
<td>.35</td>
<td>.39</td>
<td>.28</td>
<td>.10</td>
<td>.29</td>
<td>.13</td>
<td>.36</td>
<td>.30</td>
<td>.37</td>
<td>.46</td>
<td>.36</td>
<td>.22</td>
<td>.51</td>
<td>.31</td>
</tr>
<tr>
<td>U-hills</td>
<td>.55</td>
<td>.43</td>
<td>.49</td>
<td>.47</td>
<td>.33</td>
<td>.40</td>
<td>.33</td>
<td>.55</td>
<td>.56</td>
<td>.54</td>
<td>.49</td>
<td>.45</td>
<td>.60</td>
<td>.69</td>
<td>.59</td>
</tr>
<tr>
<td>V-standing</td>
<td>.02</td>
<td>.09</td>
<td>.06</td>
<td>.15</td>
<td>.08</td>
<td>.18</td>
<td>.09</td>
<td>.13</td>
<td>.09</td>
<td>.09</td>
<td>.23</td>
<td>.19</td>
<td>.06</td>
<td>.44</td>
<td>.18</td>
</tr>
<tr>
<td>W-up from kneeling</td>
<td>.52</td>
<td>.45</td>
<td>.46</td>
<td>.50</td>
<td>.42</td>
<td>.43</td>
<td>.45</td>
<td>.64</td>
<td>.62</td>
<td>.57</td>
<td>.46</td>
<td>.49</td>
<td>.89</td>
<td>.60</td>
<td>.66</td>
</tr>
<tr>
<td>X-in/out of car</td>
<td>.38</td>
<td>.50</td>
<td>.34</td>
<td>.36</td>
<td>.30</td>
<td>.37</td>
<td>.32</td>
<td>.29</td>
<td>.48</td>
<td>.48</td>
<td>.47</td>
<td>.46</td>
<td>.42</td>
<td>.62</td>
<td>.49</td>
</tr>
<tr>
<td>Y-sex</td>
<td>.07</td>
<td>.18</td>
<td>.15</td>
<td>.26</td>
<td>.17</td>
<td>.36</td>
<td>.17</td>
<td>.37</td>
<td>.29</td>
<td>.19</td>
<td>.22</td>
<td>.15</td>
<td>.30</td>
<td>.24</td>
<td>.24</td>
</tr>
<tr>
<td>Z-work duties</td>
<td>.09</td>
<td>.13</td>
<td>.25</td>
<td>.23</td>
<td>.37</td>
<td>.50</td>
<td>.34</td>
<td>.29</td>
<td>.41</td>
<td>.29</td>
<td>.18</td>
<td>.13</td>
<td>.36</td>
<td>.42</td>
<td>.34</td>
</tr>
<tr>
<td>AA-work hours</td>
<td>.21</td>
<td>.20</td>
<td>.27</td>
<td>.26</td>
<td>.34</td>
<td>.52</td>
<td>.34</td>
<td>.48</td>
<td>.46</td>
<td>.35</td>
<td>.32</td>
<td>.20</td>
<td>.43</td>
<td>.42</td>
<td>.40</td>
</tr>
<tr>
<td>AB-leisure activities</td>
<td>.48</td>
<td>.54</td>
<td>.54</td>
<td>.37</td>
<td>.48</td>
<td>.36</td>
<td>.40</td>
<td>.61</td>
<td>.66</td>
<td>.48</td>
<td>.45</td>
<td>.46</td>
<td>.52</td>
<td>.45</td>
<td>.69</td>
</tr>
<tr>
<td>AC-social activities</td>
<td>.42</td>
<td>.41</td>
<td>.44</td>
<td>.25</td>
<td>.34</td>
<td>.29</td>
<td>.46</td>
<td>.57</td>
<td>.45</td>
<td>.54</td>
<td>.50</td>
<td>.35</td>
<td>.38</td>
<td>.34</td>
<td>.47</td>
</tr>
<tr>
<td>AD-sports</td>
<td>.34</td>
<td>.32</td>
<td>.32</td>
<td>.35</td>
<td>.40</td>
<td>.51</td>
<td>.32</td>
<td>.57</td>
<td>.62</td>
<td>.44</td>
<td>.33</td>
<td>.38</td>
<td>.54</td>
<td>.56</td>
<td>.56</td>
</tr>
<tr>
<td>ITEM</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>AA</td>
<td>AB</td>
<td>AC</td>
<td>AD</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----</td>
<td>------</td>
<td>----</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>P-downstairs</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-driving</td>
<td>.22</td>
<td>1.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R- indoors</td>
<td>.54</td>
<td>.17</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-walking outdoors</td>
<td>.72</td>
<td>.22</td>
<td>.75</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-sitting</td>
<td>.31</td>
<td>.39</td>
<td>.29</td>
<td>.40</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U-hills</td>
<td>.59</td>
<td>.17</td>
<td>.54</td>
<td>.69</td>
<td>.55</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V-standing</td>
<td>.18</td>
<td>.17</td>
<td>.30</td>
<td>.25</td>
<td>.59</td>
<td>.47</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W-up from kneeling</td>
<td>.72</td>
<td>.25</td>
<td>.49</td>
<td>.62</td>
<td>.25</td>
<td>.61</td>
<td>.10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-in/out of car</td>
<td>.49</td>
<td>.28</td>
<td>.31</td>
<td>.48</td>
<td>.68</td>
<td>.60</td>
<td>.57</td>
<td>.46</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-sex</td>
<td>.24</td>
<td>.62</td>
<td>.10</td>
<td>.23</td>
<td>.30</td>
<td>.12</td>
<td>.19</td>
<td>.29</td>
<td>.33</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-work duties</td>
<td>.29</td>
<td>.29</td>
<td>.18</td>
<td>.26</td>
<td>.22</td>
<td>.36</td>
<td>.23</td>
<td>.36</td>
<td>.29</td>
<td>.29</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA-work hours</td>
<td>.40</td>
<td>.33</td>
<td>.26</td>
<td>.37</td>
<td>.33</td>
<td>.43</td>
<td>.24</td>
<td>.44</td>
<td>.32</td>
<td>.32</td>
<td>.75</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB-leisure activities</td>
<td>.57</td>
<td>.06</td>
<td>.44</td>
<td>.60</td>
<td>.37</td>
<td>.51</td>
<td>.16</td>
<td>.53</td>
<td>.44</td>
<td>.17</td>
<td>.34</td>
<td>.35</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-social activities</td>
<td>.42</td>
<td>.27</td>
<td>.34</td>
<td>.49</td>
<td>.44</td>
<td>.37</td>
<td>.00</td>
<td>.45</td>
<td>.34</td>
<td>.21</td>
<td>.38</td>
<td>.36</td>
<td>.52</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>AD-sports</td>
<td>.55</td>
<td>.35</td>
<td>.37</td>
<td>.49</td>
<td>.31</td>
<td>.58</td>
<td>.24</td>
<td>.62</td>
<td>.46</td>
<td>.36</td>
<td>.38</td>
<td>.47</td>
<td>.51</td>
<td>.29</td>
<td>1</td>
</tr>
</tbody>
</table>
TABLE 5
LOWER EXTREMITY SIX MONTHS POST-OPERATIVELY:
SUMMARY OF THE CORRELATION MATRIX

<table>
<thead>
<tr>
<th>ITEM</th>
<th>MEAN (sd) CORRELATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>standing</td>
<td>0.23 (.21)</td>
</tr>
<tr>
<td>driving</td>
<td>0.24 (.18)</td>
</tr>
<tr>
<td>sexual activities</td>
<td>0.27 (.17)</td>
</tr>
<tr>
<td>work duties</td>
<td>0.33 (.18)</td>
</tr>
<tr>
<td>sitting</td>
<td>0.37 (.17)</td>
</tr>
<tr>
<td>light household</td>
<td>0.38 (.18)</td>
</tr>
<tr>
<td>meal preparation</td>
<td>0.38 (.17)</td>
</tr>
<tr>
<td>gardening/yard work</td>
<td>0.39 (.15)</td>
</tr>
<tr>
<td>work hours</td>
<td>0.39 (.16)</td>
</tr>
<tr>
<td>social activities</td>
<td>0.40 (.16)</td>
</tr>
<tr>
<td>pants</td>
<td>0.41 (.21)</td>
</tr>
<tr>
<td>walking indoors</td>
<td>0.42 (.18)</td>
</tr>
<tr>
<td>showering</td>
<td>0.42 (.16)</td>
</tr>
<tr>
<td>socks</td>
<td>0.43 (.18)</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>0.43 (.19)</td>
</tr>
<tr>
<td>shoes</td>
<td>0.43 (.20)</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>0.44 (.15)</td>
</tr>
<tr>
<td>getting in/out of a car</td>
<td>0.45 (.15)</td>
</tr>
<tr>
<td>sports</td>
<td>0.45 (.15)</td>
</tr>
<tr>
<td>leisure activities</td>
<td>0.47 (.17)</td>
</tr>
<tr>
<td>bending</td>
<td>0.48 (.18)</td>
</tr>
<tr>
<td>kneeling</td>
<td>0.48 (.14)</td>
</tr>
<tr>
<td>bathtub</td>
<td>0.49 (.17)</td>
</tr>
<tr>
<td>hills</td>
<td>0.50 (.16)</td>
</tr>
<tr>
<td>going downstairs</td>
<td>0.51 (.19)</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>0.51 (.18)</td>
</tr>
<tr>
<td>heavy household</td>
<td>0.52 (.17)</td>
</tr>
<tr>
<td>shopping</td>
<td>0.52 (.17)</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>0.53 (.18)</td>
</tr>
<tr>
<td>going upstairs</td>
<td>0.53 (.18)</td>
</tr>
</tbody>
</table>

The matrix is 0.41, sd=0.15. Examination of the correlation matrix shows that going upstairs (mean \(r=0.53, sd=0.18\)) and walking outdoors (mean \(r=0.53, sd=0.18\)) had the highest mean item-to-item correlation and standing (mean \(r=0.23, sd=0.21\)) had the lowest mean item-to-item correlation. Driving and standing have correlations in the range of 0.30 to 0.70 with only four other items. Sexual activity has only seven of thirty items...
in this range. Driving has a mean item-to-item correlation of 0.24, sd=0.18 and sexual activity has a mean item-to-item correlation of 0.27, sd=0.17. Work duties similarly has a low proportion of items with item-to-item correlations between 0.3 and 0.7 (12/30). The mean item-to-item correlation is above Ferketich's suggested criteria of 0.4 (Ferketich, 1991) with the exception of the following items: light household activities (0.38); gardening/yard work (0.39); meal preparation (0.38); driving (0.24); sitting (0.37); standing (0.23); sexual activities (0.27); work duties (0.33); and, work hours (0.39). However, five of these items have mean correlations very close to the cut-point.

The item-to-item correlation pattern obtained from these data is different from that of the previous data (Davis, 1994). In the Master's thesis data (Davis, 1994), the mean item-to-item correlation for the matrix was 0.48, sd=0.16 and only five items, light household activities, gardening/yard work, meal preparation, sexual activities and driving, had mean item-to-item correlations below 0.40 as compared to the nine items in these data. Furthermore, the mean item-to-item correlations for the remaining twenty-one items of the current data set are uniformly lower with larger standard deviations. The low mean item-to-item correlations for driving and sexual activity in the current data are similar to previous observations where driving had a mean item-to-item correlation of 0.29, sd=0.12 and sexual activity had a mean item-to-item correlation of 0.24, sd=0.10 (Davis, 1994). The current data suggest that the inter-item relationships are not as strong as those noted in the previous study.

Cronbach's alpha for the items is 0.94 concurring with the value obtained in the Master's data set (Davis, 1994). This seems somewhat contradictory considering the lower item-to-item correlations in this data compared to the previous results. However, the consistent value of alpha reflects how the metric is influenced by large numbers of items.
Factor Analysis

The factor analysis at six months was based on n=94 as three subjects did not complete the TESS at this time point and no imputation was used for individual items. Items coded as "not applicable" were imputed as a value of "1-impossible to do" to prevent missing values. This imputation is identical to that used in the previous analysis (Davis, 1994). The mean value is often used in imputation for missing values (Ware, 1993), however, the "not applicable" response for the TESS items is an indication that the individual does not perform the activity as a part of his or her normal daily routine. Whether the response is truly a reflection of personal choice or the result of a disability preventing the performance of the activity is unknown. Consequently, the "impossible to do" response is a more conservative reflection of the individual's ability. The results of the factor analysis using this imputation for the six months post-operative data are presented.

Using an eigenvalue greater than or equal to one as the criterion for maintaining factors, the principal component analysis using the unadjusted correlation matrix of the difficulty ratings identified six unrotated factors that explained 44.68, 7.70, 6.43, 4.62, 4.05, and 3.63 percent of the variance respectively (cumulative variance of 71.01%). The unrotated factors identify a primary general factor; all items but driving and standing load on this factor. Rounding to two decimal places, driving and standing have factor loadings of 0.33 and 0.32 respectively which is just below the recommended cut-point of 0.35 (Nunnally, 1978). The fact that these items do not load on the first factor is not surprising in reviewing the correlation matrix because, as previously noted, both of these items had only four of thirty items with correlations between the range of 0.40 and 0.70, with the remaining item-to-item correlations below 0.40. Six items, driving, sitting, standing, sexual activities, work duties and work hours, load on the second factor; three items, sitting, standing and getting in and out of a car, load on the third factor; two items, driving and sexual activity, load on factor four; social activities is the only item loading on factor five; and, meal preparation and getting in and out of bed load on the final factor.
Six items, driving, sitting, standing, sexual activity, work duties and work hours all load on more than one factor.

Orthogonal rotation of the factors was undertaken in an attempt to obtain simple structure, that is to achieve a smaller number of factors and to have items correlated greater than 0.35 on only one factor. Orthogonal rotation rotates the factors so that they remain uncorrelated. The items again loaded across six factors with the factors explaining the same total proportion of variance as the unrotated factors; the number of items loading on multiple factors increased to seventeen (Table 6). Although the intent of rotation was to obtain simple structure, this rotation did not reduce the number of factors nor did it achieve individual items loading on a single factor. In fact, the complexity increased. There is no conceptual meaning to the orthogonally rotated solution.

Oblique rotation was then undertaken on the factors. Oblique rotation allows factors to be correlated with one another in an attempt to obtain simple solution. The simplest structure resulted in six factors with seven items (gardening/yard work, meal preparation, bending to pick-up objects off the floor, going upstairs, sitting, hills, and social activities) loading on multiple factors. The six factors explained 86.98 percent of the total variance. Factor rotation did not provide simple structure.

These results were somewhat contrary to the factor analysis results from the Master's thesis data set (Davis, 1994) in which all items loaded strongly on the first factor with few items loading on multiple factors. Factor rotation provided similar results to the unrotated analysis. In attempting to interpret the current factor analysis results as compared to the previous findings (Davis, 1994), a difference in the length of follow-up of the subjects in the two samples was noted. The prior work (Davis, 1994) had a mean follow-up of twenty-three months compared to the six months in the current data. To evaluate whether follow-up time had an effect, the twelve months data (when functional status is more stable in patients who have not relapsed locally or systemically) for the current study was factor analyzed. The results of this analysis were similar to that of the
TABLE 6 - PRINCIPAL COMPONENT ANALYSIS ORTHOGONAL ROTATION
AT SIX MONTHS POST-OPERATIVELY
*Bolded items load on more than one factor

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>pants</td>
<td>.31</td>
<td>-.01</td>
<td>.10</td>
<td>-.08</td>
<td>.77</td>
<td>.18</td>
<td>0.74</td>
</tr>
<tr>
<td>shoes</td>
<td>.25</td>
<td>.03</td>
<td>.14</td>
<td>.05</td>
<td>.83</td>
<td>.12</td>
<td>0.79</td>
</tr>
<tr>
<td>socks</td>
<td>.28</td>
<td>.20</td>
<td>.09</td>
<td>.00</td>
<td>.73</td>
<td>.10</td>
<td>0.67</td>
</tr>
<tr>
<td>shower</td>
<td>.40</td>
<td>.04</td>
<td>.16</td>
<td>.13</td>
<td>.22</td>
<td>.54</td>
<td>0.55</td>
</tr>
<tr>
<td>*light household</td>
<td>.30</td>
<td>.61</td>
<td>-.02</td>
<td>-.15</td>
<td>.17</td>
<td>.37</td>
<td>0.65</td>
</tr>
<tr>
<td>garden/yard</td>
<td>.44</td>
<td>.52</td>
<td>.17</td>
<td>.27</td>
<td>.17</td>
<td>-.07</td>
<td>0.60</td>
</tr>
<tr>
<td>meal preparation</td>
<td>.14</td>
<td>.55</td>
<td>.00</td>
<td>-.10</td>
<td>.18</td>
<td>.54</td>
<td>0.66</td>
</tr>
<tr>
<td>shopping</td>
<td>.47</td>
<td>.26</td>
<td>.03</td>
<td>.28</td>
<td>.43</td>
<td>.42</td>
<td>0.73</td>
</tr>
<tr>
<td>heavy household</td>
<td>.58</td>
<td>.40</td>
<td>.10</td>
<td>.10</td>
<td>.35</td>
<td>.29</td>
<td>0.82</td>
</tr>
<tr>
<td>bathtub</td>
<td>.28</td>
<td>.23</td>
<td>.11</td>
<td>.10</td>
<td>.61</td>
<td>.35</td>
<td>0.66</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>.08</td>
<td>.11</td>
<td>.28</td>
<td>.24</td>
<td>.39</td>
<td>.61</td>
<td>0.68</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>.26</td>
<td>-.06</td>
<td>.25</td>
<td>.03</td>
<td>.55</td>
<td>.42</td>
<td>0.62</td>
</tr>
<tr>
<td>kneeling</td>
<td>.76</td>
<td>.21</td>
<td>.04</td>
<td>.13</td>
<td>.28</td>
<td>.20</td>
<td>0.76</td>
</tr>
<tr>
<td>bending</td>
<td>.50</td>
<td>.26</td>
<td>.55</td>
<td>.01</td>
<td>.18</td>
<td>.21</td>
<td>0.70</td>
</tr>
<tr>
<td>upstairs</td>
<td>.51</td>
<td>.20</td>
<td>.14</td>
<td>.02</td>
<td>.46</td>
<td>.46</td>
<td>0.75</td>
</tr>
<tr>
<td>downstairs</td>
<td>.60</td>
<td>.14</td>
<td>.14</td>
<td>.14</td>
<td>.35</td>
<td>.40</td>
<td>0.71</td>
</tr>
<tr>
<td>driving</td>
<td>.11</td>
<td>.06</td>
<td>.17</td>
<td>.87</td>
<td>.00</td>
<td>.10</td>
<td>0.81</td>
</tr>
<tr>
<td>walking indoors</td>
<td>.34</td>
<td>.02</td>
<td>.24</td>
<td>.02</td>
<td>.12</td>
<td>.74</td>
<td>0.74</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>.42</td>
<td>.13</td>
<td>.21</td>
<td>.13</td>
<td>.36</td>
<td>.64</td>
<td>0.80</td>
</tr>
<tr>
<td>sitting</td>
<td>-.05</td>
<td>.12</td>
<td>.75</td>
<td>.32</td>
<td>.37</td>
<td>.10</td>
<td>0.84</td>
</tr>
<tr>
<td>hills</td>
<td>.52</td>
<td>.15</td>
<td>.57</td>
<td>-.05</td>
<td>.31</td>
<td>.26</td>
<td>0.78</td>
</tr>
<tr>
<td>standing</td>
<td>.05</td>
<td>.06</td>
<td>.89</td>
<td>.04</td>
<td>-.13</td>
<td>.11</td>
<td>0.83</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>.75</td>
<td>.19</td>
<td>.07</td>
<td>.16</td>
<td>.30</td>
<td>.28</td>
<td>0.79</td>
</tr>
<tr>
<td>getting in/out of car</td>
<td>.26</td>
<td>.14</td>
<td>.69</td>
<td>.15</td>
<td>.35</td>
<td>.13</td>
<td>0.73</td>
</tr>
<tr>
<td>sex</td>
<td>.21</td>
<td>.15</td>
<td>.09</td>
<td>.78</td>
<td>.05</td>
<td>.04</td>
<td>0.69</td>
</tr>
<tr>
<td>work duties</td>
<td>.21</td>
<td>.79</td>
<td>.21</td>
<td>.16</td>
<td>.02</td>
<td>-.01</td>
<td>0.74</td>
</tr>
<tr>
<td>work hours</td>
<td>.31</td>
<td>.67</td>
<td>.23</td>
<td>.25</td>
<td>.07</td>
<td>.05</td>
<td>0.67</td>
</tr>
<tr>
<td>leisure activities</td>
<td>.37</td>
<td>.35</td>
<td>.16</td>
<td>-.03</td>
<td>.50</td>
<td>.27</td>
<td>0.60</td>
</tr>
<tr>
<td>social activities</td>
<td>-.06</td>
<td>.46</td>
<td>.05</td>
<td>.27</td>
<td>.54</td>
<td>.36</td>
<td>0.70</td>
</tr>
<tr>
<td>sports</td>
<td>.69</td>
<td>.25</td>
<td>.23</td>
<td>.21</td>
<td>.15</td>
<td>.09</td>
<td>0.67</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>4.92</td>
<td>2.95</td>
<td>3.30</td>
<td>2.08</td>
<td>4.05</td>
<td>4.02</td>
<td></td>
</tr>
</tbody>
</table>
six months time, and again failed to provide simple structure (Appendix Four). There are seven factors and after varimax rotation, fifteen items loaded on multiple factors and three items loaded on multiple factors after oblique rotation.

The second potential explanation for the lack of simple structure is the recognized clinical heterogeneity of the extremity population. Based on the clinical expertise of one of the committee members (RSB), potentially homogeneous sub-groups of patients were identified and factor analysis was again conducted on these sub-groups. The sub-groups were soft tissue tumours, bone tumours, pelvic/hip tumours, small tumours (less than the median maximum tumour diameter of 13.4 cm), larger tumours, no wound complications, and wound complications. None of the analyses of these sub-groups provided simple data structure as shown by the summary in Table 7.

Table 7

<table>
<thead>
<tr>
<th>Sub-group</th>
<th>Number of Factors</th>
<th>Items Loading on Multiple Items Varimax Rotation</th>
<th>Oblique Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Tissue</td>
<td>8</td>
<td>12 items</td>
<td>7 items</td>
</tr>
<tr>
<td>Bone Tissue</td>
<td>6</td>
<td>17 items</td>
<td>14 items</td>
</tr>
<tr>
<td>Pelvic/hip Tumours</td>
<td>6</td>
<td>10 items</td>
<td>10 items</td>
</tr>
<tr>
<td>Small Tumours</td>
<td>7</td>
<td>13 items</td>
<td>8 items</td>
</tr>
<tr>
<td>Large Tumours</td>
<td>7</td>
<td>18 items</td>
<td>5 items</td>
</tr>
<tr>
<td>No Wound Complications</td>
<td>7</td>
<td>16 items</td>
<td>12 items</td>
</tr>
<tr>
<td>Wound Complications</td>
<td>6</td>
<td>10 items</td>
<td>12 items</td>
</tr>
</tbody>
</table>

In summary, PCA does not provide simple structure for the lower extremity TESS but rather suggests that the items of the lower extremity TESS are not inter-related and are not well explained by a single, general factor, physical function.

The factor analyses of the items suggest that there is not an underlying construct represented by a linear solution for the TESS items. Rather the items are considered to be related on the basis of a clinimetric interpretation, that is, the items were chosen on the basis of clinical sensibility including clinician and patient impressions of important items.
that reflect physical function. The clinimetric interpretation provides the justification for summatiing the item scores into a single value.

VALIDITY STUDY

Construct Validity

Construct validity was evaluated by determining whether the TESS had anticipated relationships with measures of related constructs. The TESS was anticipated to have a high correlation with the physical function subscale of the SF-36 as the two measures contain similar items. This relationship was confirmed with the Spearman's correlation coefficients of 0.88 pre-operatively, 0.80 at six weeks, 0.85 at three months, 0.89 at six months and 0.88 at twelve months post-operatively. Social and mental health scales typically have moderately positive relationships with physical health and confirmation of these relationships was expected in the current study sample. The social function subscale of the SF-36 had moderately positive correlations with the TESS and the correlation coefficients were 0.66, 0.49, 0.65, 0.65, and 0.68 at the five time points outlined above. Similarly, the mental health subscale and the TESS had moderate correlations with coefficients of 0.33, 0.33, 0.44, 0.49, and 0.43 at the time points respectively. Consequently, the expected relationships of the TESS and the physical function, social function and mental health subscales were demonstrated.

Kendall's tau was used to evaluate the relationship of the TESS and the MSTS (1987). The correlations pre-operatively, six weeks, three months, six months and twelve months post-operatively were 0.49, 0.54, 0.54, 0.56, and 0.50 respectively. The correlations of the 1993 version of the MSTS with the TESS were higher with values of 0.68, 0.81, 0.87, 0.78 and 0.89. Again, the expected relationships of the TESS and the two versions of the MSTS were demonstrated.

Table 8 presents a summary of the correlation coefficients for the five time points between all measures, that is the TESS, the MSTS 1987 and 1993 versions, and the eight
### Table 8: Range of Correlational Relationships of the TESS, MSTS (1987, 1993) and SF-36 Subscales over Repeated Measures

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TESS</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSTS (1987)</td>
<td>0.49*</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSTS (1993)</td>
<td>0.68</td>
<td>0.77</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>0.80</td>
<td>0.55</td>
<td>0.67</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>0.65*</td>
<td>-0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP</td>
<td>0.61</td>
<td>0.45</td>
<td>0.49</td>
<td>0.62</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.78</td>
<td>0.49*</td>
<td>0.71</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>0.39</td>
<td>0.34</td>
<td>0.31</td>
<td>0.43</td>
<td>0.37</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.68</td>
<td>0.55*</td>
<td>0.66</td>
<td>0.74</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>0.33</td>
<td>0.24</td>
<td>0.17</td>
<td>0.42</td>
<td>0.32</td>
<td>0.23</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.66</td>
<td>0.39*</td>
<td>0.51</td>
<td>0.62</td>
<td>0.58</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VT</td>
<td>0.40</td>
<td>0.25</td>
<td>0.34</td>
<td>0.46</td>
<td>0.35</td>
<td>0.44</td>
<td>0.43</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.37*</td>
<td>0.50</td>
<td>0.58</td>
<td>0.61</td>
<td>0.55</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF</td>
<td>0.49</td>
<td>0.25</td>
<td>0.38</td>
<td>0.51</td>
<td>0.44</td>
<td>0.44</td>
<td>0.33</td>
<td>0.41</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.68</td>
<td>0.47*</td>
<td>0.60</td>
<td>0.69</td>
<td>0.73</td>
<td>0.60</td>
<td>0.61</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.31</td>
<td>0.27</td>
<td>0.34</td>
<td>0.35</td>
<td>0.29</td>
<td>0.42</td>
<td>0.37</td>
<td>0.33</td>
<td>0.33</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.59</td>
<td>0.34*</td>
<td>0.46</td>
<td>0.64</td>
<td>0.71</td>
<td>0.58</td>
<td>0.54</td>
<td>0.50</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>0.33</td>
<td>0.30</td>
<td>0.27</td>
<td>0.40</td>
<td>0.18</td>
<td>0.33</td>
<td>0.37</td>
<td>0.59</td>
<td>0.33</td>
<td>0.33</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>0.49</td>
<td>0.33*</td>
<td>0.46</td>
<td>0.53</td>
<td>0.51</td>
<td>0.54</td>
<td>0.60</td>
<td>0.78</td>
<td>0.56</td>
<td>0.49</td>
<td></td>
</tr>
</tbody>
</table>

Correlations represent the range of r for the five time points of questionnaire completion. Correlations were calculated using Spearman's Rank Coefficient except where indicated * where the coefficient is Kendall's tau.
subscales of the SF-36. With the exception of the strong positive correlations observed between the TESS and the physical function subscale of the SF-36, the TESS has moderately positive relationships with all the remaining measures. These relationships support the construct validity of the TESS.

**Criterion Validity**

In evaluating criterion validity, the TESS scores were compared to the global rating of disability to determine if the TESS scores discriminated amongst subjects' self-ratings of overall disability. The overall F-test of the one-way analysis of variance for the analysis was significant (p<0.001) at each time point. The results are presented graphically in Figure 4 (TESS vs. disability) for each time point. The results demonstrate that the TESS scores discriminated amongst subject ratings of overall disability. The three and twelve months post-operative periods demonstrate an overlap in the 95% confidence interval for the "severely disabled" and "moderately disabled" ratings. However, the sample sizes are small for these ratings of disability.

The depth of the post-operative scores on the TESS with floor and ceiling effects of less than 3% (Figure 5) indicates that patients experience a wide range of physical disability. TESS scores at six months post-surgery range from twenty-nine to one hundred and from thirty-five to one hundred at one year post-surgery. On average, subjects with TESS scores greater than ninety rated themselves as having no disability and those with scores between sixty-five and ninety rated themselves as mildly disabled. As the number of subjects who rated themselves as severely disabled at any given time point was small (six or less subjects), it was difficult to confidently differentiate between those who considered themselves moderately versus severely disabled. However, with the exception of one subject at six months and one year post surgery, all subjects who rated themselves as severely disabled had a TESS score below thirty-five. At six months post surgery, twenty-five percent of the sample had no disability, fifty percent were mildly
disabled and the remaining twenty-five percent were moderately to severely disabled. At one year post surgery, the percentages were similar.

Figure 4: Discrimination of TESS Scores and Disability

Pre-operative TESS Score vs. Disability Rating

Six Week Post-operative TESS Score vs. Disability Rating

Disability
1=completely disabled, 2=severely disabled, 3=moderately disabled
4=mildly disabled, 5=not at all disabled
Figure 4 cont.

Three Months Post-operative TESS Score vs. Disability Rating

Disability
1=completely disabled, 2=severely disabled, 3=moderately disabled
4=mildly disabled, 5=not at all disabled

Six Months Post-operative TESS Score vs. Disability Rating

Disability
1=completely disabled, 2=severely disabled, 3=moderately disabled
4=mildly disabled, 5=not at all disabled

One Year Post-operative TESS Score vs. Disability Rating

Disability
1=completely disabled, 2=severely disabled, 3=moderately disabled
4=mildly disabled, 5=not at all disabled
The TESS scores were hypothesized to discriminate among use of ambulatory aids most successfully at the six weeks and twelve months post-operative time points. As shown in Table 9, at six weeks post-operatively the mean TESS scores for those using no ambulatory aids were highest at eighty-two, decreasing to sixty-nine for those using a single cane, and further decreasing to fifty-one for those using crutches. At twelve months, subjects using no ambulatory aids had a mean score of ninety-one, while those using a cane again had a mean score of sixty-nine and those on crutches had a score of forty-eight. The number of subjects using two canes or a walker were too small for interpreting the TESS scores in relation to these ambulatory aids.

The three and six months post-operative times were anticipated to be less discriminatory due to surgeon preference for prolonged, protected weight-bearing following some bone reconstructive procedures. This lack of discrimination was demonstrated by the similarity in the mean TESS scores for subjects using one cane and two crutches. At three months post-surgery, the mean TESS score was sixty-two for those using one cane and fifty-nine for those using two crutches. At six months the TESS scores were sixty-five and sixty-one respectively.

Comparison of TESS scores to the external criteria of disability rating and the use of ambulatory aids suggests that the TESS has criterion validity.

Criterion validity was further evaluated by testing known group validity. The anticipated difference in pre-operative TESS scores for bone versus soft tissue tumour patients was demonstrated supporting known group validity. Bone patients had a mean pre-operative score of seventy-two and soft tissue tumour patients had a pre-operative score of eighty-four (p=0.01).

A further criterion required of measures used in monitoring is detection of change over time. Clinical judgement suggests that lower extremity sarcoma patients have a predictable pattern of recovery following surgery. The predicted clinical pattern of scores over time was demonstrated for the TESS, MSTS (1987 and 1993 versions) and the
Table 9

Mean Post-operative TESS Scores
By Use of Ambulatory Aid (n=97)

<table>
<thead>
<tr>
<th>Six weeks post-operatively</th>
<th>TESS Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambulatory Aid</td>
<td>N</td>
</tr>
<tr>
<td>1=none</td>
<td>44</td>
</tr>
<tr>
<td>2=one cane</td>
<td>16</td>
</tr>
<tr>
<td>3=two canes</td>
<td>0</td>
</tr>
<tr>
<td>4=two crutches</td>
<td>30</td>
</tr>
<tr>
<td>5=walker</td>
<td>4</td>
</tr>
<tr>
<td>6=wheelchair</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Three months post-operatively</th>
<th>TESS Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=none</td>
<td>56</td>
</tr>
<tr>
<td>2=one cane</td>
<td>19</td>
</tr>
<tr>
<td>3=two canes</td>
<td>2</td>
</tr>
<tr>
<td>4=two crutches</td>
<td>14</td>
</tr>
<tr>
<td>5=walker</td>
<td>3</td>
</tr>
<tr>
<td>6=wheelchair</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Six months post-operatively</th>
<th>TESS Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=none</td>
<td>67</td>
</tr>
<tr>
<td>2=one cane</td>
<td>19</td>
</tr>
<tr>
<td>3=two canes</td>
<td>4</td>
</tr>
<tr>
<td>4=two crutches</td>
<td>4</td>
</tr>
<tr>
<td>5=walker</td>
<td>2</td>
</tr>
<tr>
<td>6=wheelchair</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Twelve months post-operatively</th>
<th>TESS Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=none</td>
<td>68</td>
</tr>
<tr>
<td>2=one cane</td>
<td>18</td>
</tr>
<tr>
<td>3=two canes</td>
<td>0</td>
</tr>
<tr>
<td>4=two crutches</td>
<td>2</td>
</tr>
<tr>
<td>5=walker</td>
<td>1</td>
</tr>
<tr>
<td>6=wheelchair</td>
<td>0</td>
</tr>
</tbody>
</table>

physical function subscale of the SF-36 (Figures 5 and 6). There is a significant drop in TESS scores at six weeks post-operatively (78.9, sd=23.1 pre-operatively to 66.5, sd=23.9 at six weeks) with a gradual increase in scores over the subsequent post-operative period evaluated at one year (73.9, sd=22.5 at three months; 80.8, sd=18.8 at six months; and 85.5, sd=15.7 at one year). The TESS score at one year post-surgery
(85.5, sd=15.7) was higher than the pre-operative score (78.9, sd=23.1). This difference was statistically significant using a paired t-test with $t=2.82, p=0.006$.

**TESS Scores Over Time**

Figure 5: TESS scores over time with 1=pre-operative, 2=6 weeks, 3=3 months, 4=6 months, 5=12 months post-operatively

Post-operatively, there is a gradual improvement in physical function as demonstrated by the increase in TESS scores over time. Of note is the change in the
distribution of the scores over the post-operative period with the range of scores narrowing and the standard deviation decreasing as a result of subjects with previously low physical function scores improving. This pattern suggests that there are subgroups of subjects who recover rapidly while others recover over a longer period. A few subjects \( n=3 \) have such low scores that they are outliers.

![SF-36 - MEAN OF EIGHT SUBSCALES diagram](image)

**Figure 6:** Mean of the eight subscales of the SF-36 over time with 1=pre-operative, 2=6 weeks, 3=3 months, 4=6 months, 5=12 months post-operatively.
This TESS pattern is in contrast to the pattern observed for the SF-36 (Figure 6). The summary scores of eight subscales of the SF-36 were averaged into a single score. The subscale to subscale inter-correlations did not vary markedly (Appendix Two) such that the SF-36 data can be represented by an average of the subscales. Figure 7 shows that the tumour sample has low scores at six weeks that gradually improve over the year following surgery and this is similar to the TESS. However, the variability in the SF-36 scores is different than the TESS in that the standard deviations remain constant over time suggesting that those with low scores do not improve. The TESS data conversely suggest that there is improvement in this more disabled group of subjects.

The MSTS (1987) scores over time show minimal change with pre-operative scores decreasing from an average of 29.9 to 24.5 at six weeks post-operatively. The scores then increase to 30.1 at one year following surgery. The standard deviations over the time period are consistent ranging from 5.7 to 7.1. The MSTS (1993) demonstrates a similar pattern to the TESS although the variability of the scores is slightly larger with standard deviations ranging from 29.4 to 20.4. While the MSTS (1993) score pattern is similar to the TESS, it should be noted that this is a clinician completed scale of limited content that has not been tested for reliability and validity.

To further investigate the post-operative pattern of TESS scores, individual subject profiles for subgroups of patients were graphed. Figures 7 to 10 show the pattern of post-operative scores for subgroups, particularly proximal thigh soft tissue tumours (n=21), distal thigh soft tissue tumours (n=16), bone tumours treated by metallic prosthesis at the knee (n=9) and bone tumours treated by curettage and bone graft at the knee (n=9). The pattern of the scores for the subjects in these subgroups shows the overall increase in scores over time. However, there are also patterns in which subjects have returned to normal or near normal physical function at six weeks post-operatively and maintained this level. Others have a much more gradual recovery of physical function, while others demonstrate deterioration in function and subsequent improvement.
From these graphs (Figures 7 through 10), it is evident that a larger proportion of patients with soft tissue tumours have high TESS scores (minimal disability) at six weeks post-operatively and maintain their function. This is in contrast bone tumour patients at six weeks post-operatively, who have TESS scores more evenly distributed throughout the zero to one hundred range and who show gradual improvement over time. Matching the curve to the individual subject, the soft tissue patients who demonstrate minimal disability are those with small soft tissue tumours who experienced no post-operative morbidity (e.g. wound complications).

Within the soft tissue subgroup a number of individual subjects are noted to have scores that decline at various time points and this is seen more clearly in Figures 7 and 8 in which the proximal thigh and distal thigh soft tissue tumour subjects are graphed separately. In Figure 7 (proximal thigh soft tissue tumour subjects), designated individual graphs of varying patterns of recovery can be related to clinical course. Curve A reflects the pattern of an individual treated by surgery and post-operative irradiation for a small proximal thigh sarcoma. He had no difficulties at six weeks post-operatively. However, between the six and three months period, he had developed an open wound that became infected and required re-admission to hospital for intravenous antibiotics and subsequent deep packing of the wound. The wound was healed at approximately five months post-operatively and he has had no further problems. The TESS scores over time reflect this clinical course, dropping significantly at the time of the wound problem and increasing to near normal levels of function once the wound was completely healed. The impact of major wound complications requiring repeated surgery for debridement and prolonged packing are also seen for Subject B. This patient had a large proximal thigh tumour treated with pre-operative irradiation and developed a major wound complication in the first week post-surgery while still in hospital. Further surgeries were required as was prolonged deep packing of the wound and long-standing morphine use for pain control. The six week TESS score is low and changes little over the next months in which the wound
Figure 7: Proximal thigh soft tissue tumour subjects with post-operative TESS scores plotted over time. 2=6 weeks, 3=3 months, 4=6 months, 5=12 months

Figure 8: Distal thigh soft tissue tumour subjects with post-operative TESS scores plotted over time. 2=6 weeks, 3=3 months, 4=6 months, 5=12 months
continues to heal. The one year TESS score is improved but is ten points below the overall group mean of 85.5.

The curves for distal thigh soft tissue tumour subjects (Figure 8) similarly present an overall pattern of individuals improving over time. However, individual curves again reflect clinical scenarios in which subjects experienced wound complications, severe irradiation reactions, or persistent neurogenic pain following resection of a sensory nerve.

The bone tumour subgroups (Figures 9 and 10) also reflect the overall pattern of gradual improvement in physical function over time. However, the patients reconstructed with a metallic prosthesis at the knee have much lower scores at six weeks, only showing large magnitude improvements in their TESS scores beyond six months. In contrast, the group treated with curettage and bone graft appear to improve more rapidly and to attain higher TESS scores at one year than those treated by prosthesis. The profile for these individuals again relates to the clinical course, as those treated by prosthesis have more aggressive tumours requiring more extensive surgery and, in most cases, chemotherapy. Curettage and bone graft reconstructions require less extensive surgery.

A further objective of the Validity Study was to develop a quantitative profile of TESS scores over time for the above sub-groups of patients. Consequently, descriptive statistics were calculated for the TESS for the subgroups of patients profiled in Figures 7 to 10. These profile scores over time for subgroups of patients can be used comparatively for further validation of the TESS.

Twenty-one subjects (mean age of 55.3 years, male:female=2:1) had soft tissue tumours of the proximal thigh. The mean TESS scores for the proximal thigh sub-group follow the anticipated pattern with a decrease in function at six weeks post-surgery and a gradual increase to near pre-operative status at one year. The TESS scores for the pre-operative and consecutive post-operative times for the proximal thigh soft tissue sarcoma subjects were: 85.7, sd=19.9; 74.7, sd=21.6; 76.3, sd=19.7; 80.6, sd=17.6; and, 83.3, sd=16.9. The sixteen individuals in the distal thigh soft tissue sub-group were similar in
Figure 9: Subjects with bone tumours at the knee reconstructed with an endoprosthesis with TESS scores plotted over time. 2=6 weeks, 3=3 months, 4=6 months, 5=12 months

Figure 10: Subjects with bone tumours at the knee treated by curettage and bone graft with TESS scores plotted over time. 2=6 weeks, 3=3 months, 4=6 months, 5=12 months
age (mean of 53.8) but there were equal numbers of males and females. Again, the pattern of scores over time followed the anticipated pattern of recovery. For the distal thigh soft tissue tumour subjects, the scores were: 80.6, sd=20.6; 69.9, sd=22.5; 73.9, sd=20.5; 78.5, sd=19.6; and, 80.1, sd=17.3, respectively.

Nine bone tumour patients were treated by curettage and bone graft for tumours around the knee. The sample had a mean age of 29.9 years and there were five males and four females. Pre-operative TESS scores were lower than those of either the proximal or distal thigh soft tissue tumour sub-groups, as would be anticipated from the results of the known group validity study, but the data over time followed the anticipated pattern with scores decreasing post-operatively and gradually increasing in the subsequent year. The mean post-operative TESS scores consecutively over time were: 65.6, sd=24.6; 82.2, 23.3; 85.3, sd=17.8; and, 90.8, sd=12.0.

The bone tumour patients treated for tumours at the knee and reconstructed with a prosthesis numbered seven. The sample included five males and two females with a mean age of 28.4 years. Each of these patients was treated with adjuvant chemotherapy (pre- and post-operatively) and since post-operative chemotherapy continues for three months post-surgery, this sub-group of patients would be anticipated to have a slower recovery. The TESS scores over time in fact reflect this with a six weeks post-operative mean TESS score of only 39.4 points. At three months, the mean TESS score was 49.4, sd=21.3, at six months 49.4, sd=21.3 and at one year 77.4, sd=12.6. The TESS scores progressively improve over time, as anticipated, but the overall improvement is less than that of bone tumour patients treated by curettage.

While the mean TESS scores over time reflect the average pattern of recovery, there appear to be differences in the pattern, rate and extent of improvement as demonstrated by the graphical representation of the individual scores in Figures 7 to 10. These differences may be related to the type and location of the tumour, the type of reconstructive procedure and whether an individual experiences treatment complications.
The purpose of the TESS is to monitor treatment over time and the measure, therefore, needs to be able to detect change in status. Effect sizes (Cohen, 1988) comparing two points in time were used to quantify the ability of the TESS to detect change in status. The correlation for the TESS between six weeks and three months post-operatively was 0.76 with a standardized effect size of 0.65. From three to six months post-operatively, the correlation was 0.80 with an effect size of 0.74 and from six to twelve months post-operatively, the correlation was 0.81 with an effect size of 0.63. These are medium effect sizes according to the criteria of Cohen (1988). The pre-operative and twelve month post-operative TESS scores had a moderate correlation with \( r = 0.41 \) and the effect size was moderate at 0.43 (Cohen, 1988).

**TABLE 10**

**Lower Extremity: Relative Efficiency of the TESS**

<table>
<thead>
<tr>
<th>TESS compared to</th>
<th>6 wks to 3 mos</th>
<th>3 to 6 mos</th>
<th>6 to 12 mos</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSTS (1987)</td>
<td>.65/.45=1.44</td>
<td>.74/.52=1.42</td>
<td>.63/.09=7.00</td>
</tr>
<tr>
<td>MSTS (1993)</td>
<td>.65/.50=1.30</td>
<td>.74/.52=1.42</td>
<td>.63/.29=2.17</td>
</tr>
<tr>
<td>Physical Function SF-36</td>
<td>.65/.64=1.01</td>
<td>.74/.40=1.85</td>
<td>.63/.28=2.25</td>
</tr>
</tbody>
</table>

The efficiency of the TESS in detecting change was compared by a ratio of effect sizes to that of the MSTS (1987 and 1993 versions) and the physical function subscale of the SF-36. As shown in Table 10 by a ratio greater than one, the TESS is at least as efficient as the MSTS (Enneking, 1987; Enneking, 1993) and the physical function subscale of the SF-36.
CHAPTER FIVE

DISCUSSION

The management of patients with extremity sarcoma includes surgery with or without the use of adjuvant chemotherapy or irradiation to achieve local and systemic disease control. Historically, surgery included amputation but in the past twenty years surgical management has evolved such that limb preservation is recommended (Rosenberg, 1982; Conference, 1985) with amputation necessary only in rare cases. In managing a tumour in the extremity, the goal of surgery is complete removal of the tumour to prevent local recurrence while preserving as much functioning tissue as possible. In soft tissue sarcoma, the use of irradiation combined with surgery is recognized as standard management and results in local disease control rates of greater than 90% (Harrison, 1993; Lindberg, 1981; Suit, 1985). Adjuvant chemotherapy has not impacted upon disease-free survival in soft tissue sarcoma (Bramwell, 1994) and systemic relapse rates remain constant. This is in contrast to some types of bone sarcomas where the use of adjuvant chemotherapy has resulted in improved survival rates (Link, 1991). The existing standards of management regarding surgery and adjuvant therapy for patients with extremity sarcoma produce optimal local and systemic disease control using available therapies. However, the morbidity associated with these management protocols, particularly related to physical disability, has received little attention.
Previous work has suggested that extremity sarcoma patients undergoing limb preservation surgery experience significant physical disability as a result of treatment (Bell, 1989; Bell, 1991; Capanna, 1991; Karasek, 1992; Roberts, 1991). However, none of these authors, despite reporting "physical function" outcomes, truly evaluated the physical disability of the sample as defined by any of the conceptual models of disability (Nagi, 1965; WHO, 1980; Verbrugge, 1994). The MSTS (1987), representing clinician ratings of impairments, was used in all the studies except those of Karasek (1992) and Roberts (1991). Karasek (1992) devised his own physician rated scale that also included a number of impairment measures. Roberts (1991) relied on patient reports of ambulatory status following reconstruction at the knee with metallic prosthesis. The current work is the first to evaluate disability over time based on a conceptual model and to quantify the prevalence and severity of disability in the lower extremity sarcoma population treated with limb preservation surgery.

Morbidity, specifically the physical disability, imparted by various treatments is of primary concern in evaluating treatment outcomes. Consequently, a measure of physical disability that is relevant to patients and clinicians and has measurement properties of reliability and validity is necessary to evaluate current practice. The goal, ultimately, is to improve the treatment of patients with sarcoma of the extremity such that the disability experienced as a result of treatment can be decreased.

**PROPERTIES OF THE TORONTO EXTREMITY SALVAGE SCORE**

The Toronto Extremity Salvage Score (TESS) was conceptualized and developed to serve as an outcome measure for monitoring individual patient's physical disability and evaluating the impact of different treatments on physical disability within the context of clinical trials. The extremity sarcoma population is heterogeneous and a disability measure was needed in order to deal with such diverse factors as the variation in age of patients, whether patients had bone or soft tissue tumours and the variety of surgical
reconstructions. My Master's thesis (Davis, 1994) was instrumental in determining the content of the TESS. A number of broad categories of physical function based on the definition of Patrick (1989) were included in the questionnaire (Davis, 1994).

The items selected into the questionnaire were based on a clinimetric strategy of clinician judgement and patient ratings of importance. However, the item-to-item correlations and factor analysis suggested that the items represented a single construct. The data for this study were based on a single point in time with a sample of patients with varying levels of physical function (Davis, 1994).

The initial developmental study of the TESS (Davis, 1994) also included the longitudinal evaluation of test-retest reliability. In accordance with the recommended criteria of coefficients in the range of 0.90 to 0.95 for using a measure with individual patients, (McHorney, 1995; Nunnally, 1978; Streiner, 1995), the intraclass correlation coefficients (range of 0.92 to 0.97) were such that the TESS has reliability suitable for individual patient use (Davis, 1994; Davis, 1996). While this preliminary work was encouraging with respect to the measurement properties of the TESS, further evaluation was required to evaluate the construct and criterion validity of the measure, particularly its ability to detect change in physical disability over time. This was the main objective of the current thesis.

In the current study, the breadth of the TESS scores (range twenty-eight to one hundred) with minimal floor (range 0.2 to 3.7% over time) and ceiling effects (0.2% at all times) suggest that the TESS is able to measure the spectrum of severity of disability. Furthermore, the TESS scores are able to discriminate among patients with bone and soft tissue tumours, use of ambulatory aids and patients' self-ratings of disability. The discriminative ability of the TESS is important as the initial conceptualization of the TESS recognized that the extremity sarcoma population is heterogeneous and a disability measure was needed to deal with the diversity of such factors as the variation in age of patients, type of tumour and the variety of surgical reconstructions.
In order to respond to the heterogeneity in the patient population, the TESS includes a comprehensive range of activities some of which do not apply to all patients i.e. patients under the age of sixteen are not normally expected to drive cars. Therefore, individuals are allowed to rate items as "not applicable". The goal in designing a measure is to reduce respondent burden while maintaining the face and content validity and reliability of the measure. The choice was to include items that might be rated as "not applicable" in order to improve the face and content validity for all individuals versus minimizing respondent burden by choosing a standard set of items. The content validity is demonstrated by the fact that every item within the TESS has some subjects reporting difficulty with the activity, even at six and twelve months post-surgery when TESS scores appear to plateau. Each of the individual items in the TESS, therefore, contributes to the quantification of disability for this sample of lower extremity sarcoma patients.

Calculation of the summary scores of the TESS with and without the "not applicable" responses produced comparable results despite individuals using different sets of items. It was also shown that the use of the "not applicable" responses made clinical sense, was consistent over time and had a low frequency for individual subjects. In view of the similarity of the scores, the value of including items with the highest "not applicable" response frequencies must be considered. In using the TESS for group comparisons, investigators may exercise the option of using a core group of questions and omit the "not applicable" response. This core group of questions could be determined by omitting the items with the highest frequencies of the "not applicable" responses, that is sexual activity, garden/yard work, driving, work duties and work hours and sports. It must, however, be remembered that the final content of the TESS was determined by over 160 patients' perceptions of what was important to include in the questionnaire. Consequently, how the use of a core set of items without the "not applicable" response impacts upon face and content validity must be considered. Content validity is the primary justification for maintaining the "not applicable" response option. For monitoring
individual patients in whom there are individual differences and preferences for performing different activities, the use of the full range of items improves the content validity of the TESS.

Although the TESS has measurement properties suited to its intended use, the internal structure of the scale, as evaluated in the current study, is unclear. In the initial study (Davis, 1994), the items had a higher correlation with each other (mean=0.48, sd=0.16) than in the current study (mean=0.41, sd=0.15). The item-to-item relationships in the current study seemed to be dependent on variables related to tumour characteristics e.g. bone versus soft tissue tumour, treatment related variables such as wound complications and time. Factor analysis did not provide simple structure to explain the relationship of the TESS items.

The different results obtained in the factor analysis between the current study and the previous one (Davis, 1994) may be related to sample differences and increased heterogeneity in the current data. There is a difference in length of follow-up, twenty-three months in the previous sample (Davis, 1994; Davis, 1996) compared to the maximum of twelve months in the current sample. In thinking about patients with extremity sarcoma, there are striking differences in a patient's first year of follow-up compared to the second year following their surgery. The first year is characterized by prolonged recovery of physical function that seems to plateau for most patients after the six months post-operative period. However, as demonstrated by evaluation of subgroups and individual patients, this is also a time of variable recovery that is dependent on the use of adjuvant therapies e.g. chemotherapy and the occurrence of treatment complications e.g. wound healing problems. The first year for the lower extremity sarcoma population as a total group is, therefore, represented by a disparate group of individual patients. This is in contrast to the second year following surgery for individuals (without tumour recurrence) in which patients are on a more stable course of routine follow-up for monitoring disease status and potential problems related to reconstructive procedures. Consequently, in the
second year following surgery, there is reduced disparity in the group. These differences within patients with extremity sarcoma in year one versus year two following surgery may account for the observed differences in the results of the previous and current factor analyses.

There were further differences in the samples of the previous study (Davis, 1994) and the current study. The sixty-six subjects included in the initial study had a large proportion of patients with tumours around the knee (57.6%) with only 34.8% at the pelvis/hip (Davis, 1994). This is in contrast to equal proportions of pelvis/hip and knee subjects in the current sample. The type of tumours in the samples were also different with an increased percentage of soft tissue tumours in the current data, 54.6 versus 40.9% in the previous sample. This difference in type of tumour also results in differences in the type of treatment (i.e. use of adjuvant therapy) and the type of reconstructions in the two samples. Specifically, most tumours classified as pelvis/hip are soft tissue tumours which tend to be large, having an increased risk of wound complications (Peat, 1994). This group of patients would, therefore, increase the number of subjects with a variable clinical course in the first year following surgery. Although there was an attempt to evaluate item-to-item relationships by PCA using more "homogeneous" subgroups within the current study, it must be recognized that these subgroup samples were very small (ranging from fifteen to thirty subjects) such that variation on any factor other than that chosen to define the group (e.g. a wound complication in a group defined as a soft tissue tumour proximal to the knee) might spuriously affect the results of the analysis. These sample differences may also contribute to the disparity in the factor analysis results.

The sample size of ninety-seven used in factor analyzing the thirty item TESS may also have influenced the PCA results. Norman (1993) suggests that sample size for factor analysis can be determined by the stability of the correlation coefficients. More commonly, five subjects per variable (Nunnally, 1978) are recommended for factor analysis. For this study, 150 subjects would be required to meet this criteria. However, the factor analysis
results of the sample in this study were so far from achieving simple structure that it is unlikely that increasing the sample by fifty cases would achieve simple structure. If simple structure were achieved, it would suggest that the additional subjects are different than the initial ninety-seven cases.

Despite the inability to identify the underlying structure of the items in the TESS, the items maintained within the measure reflect items that are important to extremity sarcoma patients in evaluating physical function. The fact that no new items were identified for addition or deletion suggests that the content of the TESS is valid for this population. This clinimetric approach to determining the content justifies the summation of the items in the TESS into a summary score. This summary score has been shown to discriminate among groups of patients and clearly demonstrates change over time.

The pattern of recovery was as anticipated with subjects having relatively normal scores pre-operatively, low scores at six weeks post-operatively (in response to surgery) and a gradual increase in recovery to one year. The magnitude of change detected, as calculated by standardized effect sizes (Cohen, 1988), was medium. This pattern of recovery, however, varied for subgroups of patients. The diversity imparted by individual patients is reflected by the differences in the range of scores and the large standard deviations about the mean TESS score. For example, the bone tumour patients reconstructed with a metallic prosthesis had lower scores from six weeks throughout the follow-up period compared to subjects with soft tissue tumours of the distal thigh. Specifically, the distal thigh soft tissue tumour subjects had a mean score of 69.9 with a range of seven to one hundred at six weeks post-operative and a large standard deviation of 22.5. In contrast, the group reconstructed with a metallic prosthesis had a mean score of 39.4 ranging from twenty-three to one hundred with a much smaller standard deviation of 14.8. Furthermore, individual patient's TESS scores changed in response to meaningful clinical predictors e.g. treatment complications such as wound dehiscence.
Understanding this variability in outcome is also important when educating patients about their projected course of recovery over time.

**COMPARING THE TESS TO OTHER MEASURES**

The primary goal in choosing an outcome measure in clinical trials is to choose a measure with relevant content and with the highest reliability and sensitivity to change in order to minimize the number of subjects required to conduct the study. Based on criteria of reliability and validity, the existing generic (SF-36) and disease-specific measures (MSTS, 1987; MSTS, 1993) were considered as outcomes for the extremity sarcoma population.

The construct validity of the TESS was demonstrated by its relationships to the two versions of the Musculoskeletal Tumor Society Rating Scale (Enneking, 1987; Enneking, 1993) and the physical function, social function and mental health scores of the SF-36 (Ware, 1993). The TESS was moderately correlated (range 0.49 to 0.56) with the 1987 version of the MSTS as anticipated and had correlations with the 1993 version in the range of 0.60 to 0.89. These differences in the strength of the relationship of the TESS to the two versions of the MSTS were justified on the basis of the difference in content of the two measures (Enneking, 1987; Enneking, 1993). If this argument is true, it would be expected that the two versions of the MSTS would have a moderately positive relationship. However, the correlation coefficients for the two versions of the MSTS were high, in the range of 0.77 to 0.88. In reviewing the content of the two versions of the MSTS, there are three items common to both scales: pain; overall functional ability; and, emotional acceptance of the surgical procedure. The effect of these identical items in scales with seven and six items respectively is to increase the positive correlation between the 1987 and 1993 version of the MSTS (Enneking, 1987; Enneking, 1993). This accounts for the higher than anticipated correlation between the two versions of the MSTS.
The correlations of the TESS with the MSTS (1993) and physical function subscale of the SF-36 are in a range that suggests the measures are evaluating the same construct. That being the case, one must question the value of the TESS, particularly considering that it is a longer questionnaire. The MSTS (Enneking, 1993) is a clinician completed measure with untested measurement properties. While the reliability and validity of the measure could be tested, it suffers from a fundamental flaw according to today's standards in that it evaluates patient status based on clinician ratings rather than the patient's own perceptions of his/her function (Geigle, 1990; Rineberg, 1990). The physical function subscale of the SF-36 contains ten items, many of which are also included in the TESS. However, the reliability coefficients (intraclass correlation coefficient for test-retest reliability of 0.81 and 0.91 and internal consistency reliability in the range of 0.88 to 0.94) for the physical function subscale of the SF-36 (Ware, 1993; Jenkinson, 1996) are sufficient only for group comparisons. By comparison, the TESS has reliability such that is can be used for individuals with internal consistency reliability of 0.94 and test-retest reliability coefficients in the range of 0.92 to 0.97 (Davis, 1994; Davis, 1996). A primary goal of a measure for use with the extremity sarcoma population is that it be useful for monitoring patients over time. The SF-36 physical function subscale cannot serve this purpose.

The marketing of the SF-36 has been world wide (Aaronson, 1992) such that the use of this measure in clinical trials and, more recently, in patient monitoring is highly recommended (Ware personal communication, 1996). In fact, the exclusion of the SF-36 as an outcome measure in studies with health status end points almost requires justification. This study provided the opportunity to evaluate the SF-36 as an outcome measure for the extremity sarcoma population. As noted above, the SF-36 does not meet reliability standards for individual patient use but this does not provide justification for excluding the SF-36 in group comparisons. There are, however, more critical problems with the use of the SF-36 in the extremity sarcoma population that limit its use in groups.
The meaning of the SF-36 data is unclear as the underlying structure of the eight subscales as described by Ware (1993) and the two factor structure (Ware, 1994) was not maintained in this data. The breakdown of the SF-36 scale structure may be due to the differences in the populations on which the measure was developed and the sarcoma population. The Medical Outcomes Study (MOS), from which the SF-36 was developed, included a sample of 22,000 from the general population in six American cities. A subset of 3,000 individuals with targeted medical conditions, specifically, diabetes mellitus, hypertension, heart disease and depression (McHorney, 1994) was followed longitudinally. The general population and the targeted disease populations included in the MOS do not typically have disability in activities of self-care or mobility. In contrast, the sarcoma patients in this study were in the acute phases of treatment and recovery and physical disability is a prominent characteristic of the group. The data reported by Ware (1993) from the US population that is most closely related to the physical disability experienced by the sarcoma population included a subgroup of 481 individuals with back pain or sciatica in the prior six months, osteoarthritis (n=175) and 341 subjects with "musculoskeletal complaints" exclusive of diagnosed osteoarthritis or rheumatoid arthritis. These subgroups of patients, however, also differ from the extremity sarcoma sample in the current study in the chronicity of their disease and the tendency of these diagnoses to be limited to a specific anatomical region. The data from the current study suggest that the SF-36 may not have measurement properties that generalize to a clinical sample, such as the extremity sarcoma population, in active stages of recovery.

The content of the SF-36 may also contribute to the persistently large standard deviations and low scores observed for more disabled subjects. The physical function subscale contains ten items, five of which are multi-faceted: 1) vigorous activities, such as running, lifting heavy objects, participating in strenuous sports; 2) moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf; 3) lifting or carrying groceries; 4) bending, kneeling, or stooping; and, 5) bathing or dressing oneself.
These multi-faceted items are impossible to interpret as it is unclear whether the individual is responding based on his or her ability averaged across the items, or based on a particular activity. This use of multi-faceted items obscures the scores and makes it difficult to detect change. Subscales with a small number of items, e.g. the bodily pain and social function subscales and the standardization of scores out of one hundred in subscales with few items and large standard deviations similarly make it difficult to detect change.

The use of the SF-36 in the sarcoma population should be restricted to description at a single point in time and it is recommended that the summary scores for the eight subscales be used. Based on the current data, an average of the subscales may also be used. However, a prudent researcher should evaluate the correlation for the subscales in the sample prior to proceeding with averaging. There are no data for the sarcoma population supporting the use of the Physical and Mental Health Summary Scores. Furthermore, the pattern of the SF-36 data over time suggests that the SF-36 would be a poor choice as a primary outcome measure for evaluating treatment effectiveness for the lower extremity sarcoma population.

The TESS has superior reliability and validity compared to the MSTS (1987 and 1993 versions) and the SF-36, making it the instrument of choice in measuring physical disability in the extremity sarcoma population.

THE TESS AS A MEASURE OF DISABILITY

Impairment measures related to a specific joint or muscle group are too specific to serve as an outcome measure in clinical trials and for monitoring day to day activity for the extremity sarcoma population. Although a tumour may be situated in a specific anatomical location, e.g. the distal femur, the extent of surgery and the side-effects of adjuvant therapy such as irradiation often result in stiffness, weakness and swelling in the entire extremity. The rehabilitation goal is maximization of function of the entire extremity.
Disability was chosen as the appropriate conceptual level of measurement for the extremity population. Conceptually, the TESS was placed within the context of the ICIDH (WHO, 1980) definition of disability and the items included in the TESS are represented by the ICIDH disability codes (WHO, 1980).

The items included in the ICIDH disability codes (WHO, 1980) range from simple e.g. bathing, to complex tasks e.g. household tasks. Chamie (1990), in writing about the disability classification of the ICIDH (WHO, 1980), conceptualized these simple and complex tasks as functional limitations, activities of daily living (ADLs) and instrumental activities of daily living (IADLs). Functional limitations include personal activities that do not have strong culturally-defined objectives such as bending and transfers. Activity restrictions refer to purposeful activities and are subdivided into ADLs and IADLs. ADLs include basic biological functions such as bathing, dressing and eating (Katz, 1963) whereas IADLs are more complex tasks and often include activities requiring environmental interaction e.g. preparing meals and shopping (Lawton, 1969). The rehabilitation literature has widely adopted the terms activities of daily living (ADL) and instrumental activities of daily living (IADL) or performance activities of daily living to represent the concepts of simple and complex tasks (Asberg, 1989; Kempen, 1990; McDowell, 1996; Patrick, 1993; Spector, 1987; Velozo, 1995). Unlike Chamie (1990), however, these authors include "functional limitations" under ADLs.

Functional limitations, as defined by Chamie (1990) are included in disability as they are abilities beyond the organ level. Chamie (1990) separates functional limitations from ADLs in her categorization within disability based on ADLs representing purposeful activities. This distinction of functional limitations from ADL seems artificial. Although an activity such as a transfer might be practiced in a rehabilitation setting, the intent is that the skill will be generalized to the day to day real world where the transfer is performed as a component of a purposeful activity. Individuals do not perform these activities, defined as functional limitations by Chamie (1990), without the intent of accomplishing a task. It is,
therefore, reasonable to include "functional limitations" within the broader category of ADL.

Complex tasks or IADLs are within the ICIDH definition of disability. The complex tasks may involve performance of a daily activity that includes interaction with the environment but they are not handicaps as there is no valuation by self or peer (WHO, 1980) inherent in an individual's rating of the difficulty experienced while performing the activity.

The TESS includes both ADL and IADL items. ADL items included in the TESS are: putting on pants, putting on socks and shoes (dressing); showering (bathing); and, rising from a chair, getting in and out of bed (transfers). IADLs are represented in the TESS by such items as meal preparation, shopping and light and heavy household activities. The conceptualization of activities as ADLs or IADLs was not, however, the basis for choosing items for the TESS. Physical function was defined according to Patrick (1993). In reviewing the definition and examining the TESS items, simple and complex tasks are included within both the definition and measure. The simple tasks are elemental activities requiring basic body movements and are represented within the TESS by such activities as transfers e.g. getting out of bed and rising from a chair. More complex tasks include simple tasks, often in combination, to allow completion of the more complex activity. For example, one needs to be able to walk and potentially manoeuvre stairs, in order to go shopping, a more complex task.

The concept of hierarchy in ADLs was originated by Katz (1963). He developed a scale to evaluate the effects of treatment of the elderly based on the "primary biological functions" of independence in bathing, dressing, using the toilet, moving around the house and eating. Katz (1963) described how environmental factors may artificially produce scores reflecting dependence in ADLs for hospitalized or institutionalized patients. Institutions may have mandatory supervision rules for patients, e.g. while showering, that result in a dependence rating irrespective of whether the patient is capable
of independently performing the task. To prevent this artificial lowering of scores, Katz (1963) suggested that the individual be observed, even within the institutionalized setting, performing the task in order to determine the true dependence rating. Dependence in ADLs represents a severe level of disability and Katz (1976) noted that there seemed to be particular patterns to the loss of functional skills. More complex skills e.g. bathing skills are lost first and less complex skills e.g. feeding are lost later.

In her work with the elderly, Lawton (1969) characterized IADLs as hierarchical. She described human behaviour as comprised of increasingly complex tasks ranging from physical self-maintenance to IADLs to motivation and social interaction.

Using data from the Duke Older Americans Resources and Services instrument (OARS), Spector (1987) found a hierarchical relationship spanning ADL and IADL items. IADL items such as shopping and transportation were the most demanding followed by ADL items as ordered by Katz (1963), that is, bathing the most demanding to feeding the least demanding. Although Chamie (1990) does not specifically address the issue of hierarchy of activities in her discussion of disability, her inclusion of ADL and IADL activities suggests that the hierarchical model may apply to the ICIDH (WHO, 1980) definition of disability.

The factor analysis of the current data did not identify distinct dimensions of physical disability within the TESS and this may have occurred because physical disability, as conceptualized in this thesis, has a hierarchical structure. If the concepts of ADL and IADL are hierarchical when applied to the sarcoma population, there should be a rank order of difficulty inherent in the items such that the ability to complete a given IADL task means that all less difficult ADL tasks can be done by the individual. The use of factor analysis in the current study was intended to evaluate how the items within the TESS were related such that physical disability for the extremity sarcoma population could be related to an underlying model, which in the case of factor analysis, is a linear model. The conceptualization of physical function based on Patrick's (1993) definition provided a
simplistic view of physical function in which the components had no defined relationship. Consequently, there was no conceptual basis to this linear model. In contrast, redefining physical function as measured by the TESS based on ADL and IADL provides a conceptual model, which may be hierarchical, for evaluating disability within the ICIDH disablement model (WHO, 1980). Consequently, the item relationships of the TESS might be better informed by methods based on a hierarchical model using item response theory.

ADVANTAGES AND LIMITATIONS OF THE THESIS

The incorporation of patients' perceptions of important activities related to their physical function was instrumental in developing a measure that has content validity for the lower extremity sarcoma population. Furthermore, these methods resulted in the inclusion of a range of activities, simple through more complex in physical demand, such that a measure responsive to clinically meaningful events was developed.

Currently, the TESS must be considered a disease-specific measure of physical disability for patients undergoing limb preservation for lower extremity sarcoma. It is suitable for answering questions about how different treatments affect disability and for monitoring changes in disability for groups and individuals. Furthermore, it quantifies physical disability for this population such that factors e.g. treatment complications and tumour size that might be predictive of physical function can be evaluated. If the goal of implementing an outcome is to measure physical disability in the extremity sarcoma population, the TESS has a conceptual framework and measurement properties suitable for this purpose.

The current work has focused on the lower extremity sarcoma population as a total sample. Few data were presented for subgroups of subjects due to the small numbers of subjects in the more homogeneous groups. However, based on the impact of the sample heterogeneity on the results in this study and the differences in the recovery potential of
subgroups, further work with larger samples of homogeneous groups of patients is required. This is particularly important for evaluating change over time so that accurate profiles of patients' recovery can be derived and the effects of complications and different treatment interventions on physical disability can be evaluated.

As a final comment related to the ability of the TESS to detect change in status, attention needs to be turned to recommendations for calculating sample size when using the TESS as a primary outcome measure. Most clinical trials are based on post-intervention differences in groups at a single point in time and the independent effect size is calculated by the mean difference divided by the pooled variance (Cohen, 1988). However, when calculating effect sizes for paired data, the independent effect size needs to be adjusted by the correlation coefficient (Cohen, 1988). In recalculating the paired effect size and power based on the six and twelve month post-operative data of the current study, the paired effect size is 0.63 (measures correlated at 0.81, based on means of 80.8, sd=18.8 and 85.5, sd=15.7 at six and twelve months respectively). The effect size was smaller in the pilot sample at 0.43 (measures correlated at 0.33, based on means of 73.6, sd=16.2 and 79.2, sd=15.6 at six and twelve months respectively). In comparing the current data to that upon which the initial power calculations were made, it appears that the increased effect size in the current study results from the large increase in the correlation coefficient. The effect of this higher correlation is to decrease the standard deviation of the difference between the TESS scores at six and twelve months such that the effect size is increased. The reason for this difference in the correlation coefficients is unclear but emphasizes the importance of the correlation coefficient in calculating sample size and power for paired data.

The pilot sample on which initial power calculations were made had an effect size of 0.43 detectable at ninety percent power. In recalculating the power of the current study, there is a significant increase in power to greater than ninety-nine percent. While this information would not have changed the sample size used in this study as a primary goal
was to represent the clinical sample using 100 subjects, this knowledge may be applied to future studies. If the power were maintained at eighty percent with alpha=0.05, a total of 74 subjects would be required to detect an effect size of 0.63. Therefore, the net effect is that paired data correlated at greater than 0.50 requires a reduced sample size for given power, alpha and change score than when the correlation is below 0.50. The differences observed in the correlation coefficient suggests that basing sample size and power calculations on pilot data with small numbers of extremity sarcoma patients may over estimate the required number of subjects. Furthermore, homogenous subgroups of patients who are expected to have highly correlated TESS scores on repeated measures may require even smaller sample sizes.

FUTURE DIRECTIONS AND CONCLUSIONS

The TESS has measurement properties such that it is ready to be used as a primary outcome measure in a clinical trial in the extremity sarcoma population. Whether the measure will prove to have reliability and validity such that it can serve as an outcome in other patient populations e.g. trauma, hip fracture and total hip revision remains to be evaluated.

From a conceptual perspective, the development of the TESS provides the potential to empirically test existing disablement models in relation to the sarcoma population. If the conceptualization of physical disability as hierarchical can be demonstrated by item response theory, our understanding of the severity of physical disability may be greatly enhanced. Knowing the specific item on the physical disability continuum, defined by item response theory, at which an individual starts to experience performance difficulty provides information about the individual's performance on all remaining items in the scale. Items lower than the specified item in the continuum would present no difficulty and all items above the specified item would be difficult for the individual. Theoretically, this would provide a more parsimonious measure in which the
severity of an individual's disability could be quantified by a few well chosen items. Furthermore, if people recover abilities in a given order, the items in the measure become generalizable across disease entities.

Specific to the extremity sarcoma population, the evaluation of the relationship of impairment and disability may provide the potential to develop a model predictive of the severity of disability. Surgery for extremity sarcoma requires the removal of functioning muscle and/or bone in varying quantities depending on the type of tumour and its anatomical location. This results in a complex of impairments, for example, a soft tissue tumour of the proximal thigh might result in excision of the hip flexors and a portion of the quadriceps such that hip flexion is limited as is hip flexor and knee extensor strength. How this complex of impairments relates to disability as compared to impairments of only loss of motion and strength in hip flexion is unknown. The TESS provides quantification of disability to which complexes of impairments might be related. Determining the relationship between the impairments created by the excision of specific anatomical tissues and disability provides the opportunity to develop a predictive model for outcome in extremity sarcoma patients.

It is through the development of such predictive models and the evaluation of innovative treatments within the context of clinical trials that the goal of diminishing the disability experienced by this patient population will be achieved. The existence of a reliable and valid measure of physical disability is the first step toward achieving this goal.
APPENDICES
APPENDIX ONE

THE TESS WITHIN THE CONTEXT OF THE DISABILITY MODELS
**The Toronto Extremity Salvage Score:**
**A Measure of Disability**

**Introduction**

Three major models of disability have been developed over the past three decades: that of Nagi (1965, 1991); the World Health Organization (1980); and, Verbrugge (1994). These models evolved from the recognition that many diseases result in long-term sequelae of varying severity that requires description and classification beyond the disease label. Each of the three models has similarities and differences and the following prose reviews the models, discusses their similarities and differences, and suggests points for clarification. As a final component, the Toronto Extremity Salvage Score (TESS) is discussed within the context of the World Health Organization (WHO) model.

**Historical Background**

Disability models are a set of interrelated concepts that allow description and investigation of the consequences of disease (Nagi, 1965; WHO, 1980). The classical medical model of illness/concept of disease consists of a framework in which there is progression from etiology to pathology to manifestations in signs and symptoms. Labeling by disease does not indicate the severity and impact of the disease for the individual. The disability models of Nagi (1965, 1991), WHO (1980) and Verbrugge (1994) provide the extension to this medical model of disease. The goal of the disability models was to develop a system that had a medical link (health context) but that also provided a link to the larger social problem of disability.

The necessity of a model for disability was driven by rehabilitation, chronic disease interest groups, compensation and insurance companies (Albrecht, 1992; Nagi,
The high prevalence of chronic illness and the exorbitant amounts of money paid in compensation to injured workers and by insurance companies have been instrumental in raising the question of the validity of disability claims that are based upon diagnosis and impairment descriptors. The questionable validity of diagnosis and impairment measures in determining disability arises due to the fact that severity is not an intrinsic component of all disease e.g. different individuals all with osteoarthritis may have disease manifested over the spectrum of severity. Also, impairment does not necessarily result in disability (WHO, 1980).

Nagi (1965) was the first to formulate a disability model and the WHO (1980) and Verbrugge (1994) subsequently published models. The Nagi and Verbrugge models are conceptual frameworks while the WHO model was formulated to create a taxonomic classification that was compatible with the International Classification of Diseases (ICD). The authors of all three models agree that the disability literature traditionally has been plagued by a lack of conceptual clarity and terminology across disciplines (Nagi, 1965; WHO, 1980; Verbrugge, 1994)). The models are an attempt to provide conceptual clarity that will allow hypothesis testing and ultimately a better understanding of disability.

Existing Models

In the mid 1960's, Nagi (1965) presented his disability model. The model was developed with the intent of providing conceptual clarity such that the phenomenon of disability could be described and the variance in its occurrence could be explained. The basic framework of the model is presented in Figure 1.

The first component of the model is active pathology that results from infection, metabolic imbalances, traumatic injury or other etiology (Nagi, 1965). Nagi (1965) states that this step reflects two components, that of disease disturbing normal processes and simultaneously the attempt of the organ to restore normal function. There is an inherent problem in the use of active as Nagi (1965) himself says that the pathological state may be
in remission or resolved while subsequent components of the model may be present. Intuitively, the pathology, therefore, need not be active. Similarly, the mapping of the human genome and the ability of current technology to detect genetic defects potentially extends Nagi’s concept to a necessary lower unit level of function than he initially perceived for the model. However, the inclusion of "other etiology" allows for this extension. As a final point the use of etiology and direct reference to pathology resulting from named entities is limiting as the underlying pathology is often unknown. A more general definition describing pathology as inclusive of all root causes would be more reflective of current knowledge.

The second level of Nagi’s model is impairment, defined as anatomical and/or physiological abnormalities and losses (Nagi, 1965). In his 1991 writings, Nagi uses a slightly different definition in that the losses and abnormalities are described explicitly as being of an anatomical, physiological, mental or emotional nature. This latter definition was adopted to clarify the application of the model to mental illness. An impairment cannot occur without pathology, although this pathology may be resolved or the result of a congenital abnormality. Not all impairments result in functional limitations. Impairments occur at the tissue level (Nagi, 1991).

*Functional limitations* are the limitations "set on by impairments in an individual’s ability to perform tasks and obligations of his usual roles and normal activities" (Nagi, 1965, p. 314) and form the third level of the model. Consequently, these limitations are not only dependent upon the level of severity of the impairment but also on role demands and expectations. Functional limitations are manifested at the organ level as a whole. Nagi (1965) says that functional limitations are not dependent on the existence of an impairment and cites the example of technological unemployment. This example diverges from the original conceptualization in which the model was based in a health context (Nagi, 1991). Unemployment driven purely from a societal/environmental context without an underlying
impairment does not produce a functional limitation within the health context of Nagi's disability model.

The final level of the Nagi model (Nagi, 1965) is disability. Disabilities occur at the level of social functioning and are limitations in or the inability to perform socially defined tasks and roles expected of an individual within a sociocultural and physical environment. These tasks and roles are subsumed in family/interpersonal relations; work, employment and other economic pursuits; and, education, recreation and self-care activities. Disabilities follow from functional limitations but, as noted previously, expectations and role demands (a part of the conceptualization of disability) influence the functional limitations experienced. However, not all functional limitations result in disability as disabilities are behaviours that evolve as a result of long-term impairments and functional limitations.

The use of "role" in both the functional limitations and disability categories of Nagi presents a problem of clarity. "Role" and "obligations" (used in functional limitations) imply an expected level of performance and the expectation level can be internally or externally determined, although no clarification is provided for "roles" and "obligations" as defined in this category. In the disability category, however, the "roles" and "tasks" are those "expected within a sociocultural and physical environment". The difference between "roles" in the functional limitations and disability categories is unclear. If one considered that Nagi meant functional limitations to refer to more elemental levels of purposeful activities whereas disability refers to the actions within the sociocultural and physical environment, inclusion of self-care within disability disputes this interpretation. The boundary between functional limitations and disability is unclear in the Nagi model.

The International Classification of Impairments, Disabilities, and Handicaps (ICIDH) model (WHO, 1980) was conceptualized for reasons similar to those of Nagi (1965). It does, however, have the additional component in that it was designed to provide a taxonomic framework that could be added to the ICD. The ICIDH classification
"provides a conceptual framework relevant to the long-term consequences of disease, injuries, disorders and is applicable to health care for early identification and prevention and to the mitigation of environmental and societal barriers." (WHO, 1980 p. 2) As will be described below it has similarities and differences with the Nagi model (Nagi, 1965; Nagi, 1991).

The ICIDH model begins with the conceptualization of disease. Disease occurs when an abnormality occurs within the individual, either at birth or acquired later in life. The model refers to etiology as causal circumstances that give rise to the pathology, change the structure or functioning of the body. The underlying causes may be sub-clinical or unmanifested or alternatively, manifestation may occur in signs and symptoms without recognition of underlying pathology. This definition allows extension of the definition of disease to the basic level of the human genome.

The next level of the ICIDH model is impairment that results from the "exteriorization" of the disease and, in the context of health, is any loss or abnormality of psychological, physiological or anatomical structure or function. Impairments are independent of the underlying etiology in that how the altered state developed does not determine the impairment. The loss or abnormality does not need to be perceived by the individual. This is not a contradiction to the concept of "exteriorization" of pathology as reflected by sub-clinical entities. Impairments may be temporary or permanent and reflect the individual function of parts of the body. The 1980 version of the ICIDH has subsumed functional limitations under impairments. The ICIDH model is very similar to that of Nagi (1965) with the exception that functional limitations have been included within the impairment designation in the ICIDH. Consequently, the division lines between the two models may be different with the ICIDH definition of impairment encompassing a broader concept with the inclusion of functional limitations.

Disability is any restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being (WHO, 1980). Disability results
from impairment and is restriction or limited ability to perform complex activities of the body as a whole. These activities are represented by skills, tasks and behaviours. Disability overlaps with Nagi's concept of functional limitations (Nagi, 1965; Nagi, 1991) with the common thread that they represent the function of an organ as a whole. However, it also overlaps with Nagi's disability in that Nagi (1965, 1991) includes not only role functions but self-care activities within his final level. The ICIDH includes self-care in the second to last level (also called disability in the WHO model).

The final component of the ICIDH model is handicap. Handicap results from impairment and disability and is the limitation or prevention of a role that is normal (depending on age, gender and social and cultural factors) for the individual (WHO, 1980). Handicap includes six key survival roles: orientation; physical independence; mobility; occupation; social integration; and economic self-sufficiency plus any other possibilities that might lead to disadvantage. There is a value judgement inherent in handicap in that there is judgement by self or peer group and that value judgement is in relation to cultural and societal norms. Nagi (1965, 1991) refers to similar valuation in his discussion of influences on disability. There is inherent value judgement in determining potential for rehabilitation as individual motivation is a significant variable and similarly there is valuation as peers define and react to the individual's situation. The major difference in the Nagi and WHO models in their final level is in the WHO (1980) reference to handicap as a disadvantage to the individual representing social and environmental consequences. The second difference is that Nagi (1965, 1991) includes role functions in both functional limitations and disability whereas the WHO (1980) model restricts this component to handicap. Consequently, there is overlap of the categories between the Nagi and WHO models as shown in Figures 1 and 2.

Verbrugge (1994) presents a third model using the same category labels as Nagi (1965). The first category, *pathology* is similar to both that of Nagi (1965) and the ICIDH
It is all encompassing and also recognizes that the underlying pathology may be unrecognized, evidenced only in signs and symptoms.

Impairment by Verbrugge's definition includes dysfunction and abnormalities in specific body systems that have consequences for physical, mental and social function (Verbrugge, 1994). The impairments may occur at the primary site or secondary site as a consequence of a disease. An example would be diabetes in which the primary impairment occurs in the pancreas with secondary impairments occurring in the vascular and visual systems. Measurement of impairments within Verbrugge's definition are made by clinical, laboratory or radiographic examinations (Verbrugge, 1994).

Verbrugge's functional limitations are defined as restrictions performing fundamental physical and mental actions (tasks) used in daily life by one's peers (Verbrugge, 1994). The actions are defined as generic tasks that might be used in many different situations. However, as a component of this category, Verbrugge (1994) includes many "tasks" such as "discrete motions and strengths, vision, hearing". This is contrary to the model of Nagi (1965, 1991) and the ICIDH (WHO, 1980) in which these items would be included within impairments. Consequently, the lower boundary of Verbrugge's (1994) functional limitations is shifted downward to overlap with impairment as defined by the Nagi (1965, 1991) and WHO (1980) model.

Verbrugge's final level is disability. Disability is defined as difficulty experienced doing activities in any domain of life and are distinguished from functional limitations in that they are roles not tasks (Verbrugge, 1994). The domains of life are all inclusive, ranging from employment to leisure to hygiene to sleep to activities of daily living. Disability within this definition differs from the final level of the WHO (1980) - handicap - in the domains of life that are included. The WHO model includes hygiene, sleep at the lower level (disability in the WHO model). The role inclusion in the functional limitation definition of Nagi (1965, 1991) also overlaps with Verbrugge's concept of disability. The
Nagi (1965, 1991), WHO (1980), and Verbrugge (1994) models are similar, however, in their inclusion of a social context at this final level.

As a final comment regarding disability, Verbrugge (1994) describes disability as the gap between person capability and the activity's demand. This discrepancy can be lessened either by increasing capability (e.g. through rehabilitation) or lessening the demand e.g. by environmental modifications, thereby reducing frequency. While these mechanisms may reduce demand, they do not eliminate the disability as defined by Verbrugge (1994) as the performance of activities at this level are based upon performance "typical for the age-sex group". The difference implicit in the need for modification of demand, surely must mean that there is disability.

Verbrugge's model (Verbrugge, 1994) diverges from the WHO (1980) and Nagi (1965, 1991) models in the introduction of personal assistance, external supports and aids as "interventions". Interventions also include medical care and rehabilitation and are intended to prevent, slow or reverse the disablement process. The ICIDH model (WHO, 1980), however, codes assistance within the dependence component of handicap. Verbrugge's schematic of the model (Verbrugge, 1994 p. 4) indicates that these interventions are brought to bear at the functional limitation level.

The structure of the ICIDH model differs from that of Nagi (1965, 1991) and Verbrugge (1994) in that there is also a classification system. Impairment and disability have been categorized and numbers assigned so that the digits can be used as a coding extension of the ICD codes. The intent is to identify specific impairments and disabilities associated with the ICD and to allow simplification of the impairments and disabilities to numerical form. The coding for disability is such that disability represents a "failure in accomplishment", and gradation rather than a threshold is necessary for scaling; an optional digit is allowed for this grading of severity. Similarly, an optional digit is allowed to grade for potential improvements (WHO, 1980).
The rating and coding in the classification of handicap has evolved from a slightly different perspective than impairment and disability (WHO, 1980). This is a result of the definition of handicap in which handicap is viewed as a disadvantage due to circumstances that disabled persons might encounter. These circumstances might result in the individual being disadvantaged in comparison to peers when judged by societal norms. Each of the six survival roles and other handicaps are scored from zero to nine, with increasing values indicating more severe handicap as per the definitions at each scale level. The definitions for scoring are based upon integration of performance of activities, the environment and the circumstances in which the activity must be executed. This integrative component to the rating diverges from the piecemeal, item by item rating under broad categories and optional ranking of severity and rehabilitation/recovery potential in disability (WHO, 1980).

Neither Nagi (1965, 1991) nor Verbrugge (1994) have attempted to develop their models beyond the conceptual definitions and consequently their models have not fallen to the criticisms of the WHO model (1980). Much of Nagi (1980) and Verbrugge's (1994) criticism of the ICIDH model (WHO, 1980) relates to the inclusion of similar components in the disability and handicap levels. However, the important difference to note about the ICIDH handicap level and the disability levels of Nagi and Verbrugge is that the handicap occurs only in circumstances in which the disabled individual may find her/himself disadvantaged in comparison to peers when judged by societal norms. Disability for Nagi and Verbrugge (the final level that might be viewed on level with the WHO's handicap) is defined to include more basic components such as hygiene and activities of daily living and are represented as difficulties in performing the activities (Nagi, 1965; Nagi, 1991; Verbrugge, 1994), not as placing the individual at a disadvantage in relation to peers (WHO, 1980).

All three models (Nagi, 1965; Nagi, 1991; WHO, 1980; Verbrugge, 1994) have definitional problems inherent in attempting to conceptualize abstract constructs. The three
models of disability have similarities and differences as outlined above. There are blurred boundaries within the categories of a given model e.g. Nagi's functional limitations and disability, and there are between model boundary differences in the four categories used for classification in the models (see Figures 1 and 2). Verbrugge's discussion of disability as a gap between capability and activity demands contradicts her definition of disability. Conceptually, there are differences in the final level of the models in that the ICIDH (WHO, 1980) handicap level is different than the disability level defined by Nagi (1965, 1991) and Verbrugge (1994). The result is that although each of the models is thought provoking and adds information to the concept of disability, none of the models has succeeded in providing conceptual clarity nor a model that facilitates hypothesis testing.

Where does the TESS fit into the disability model?

The TESS was conceptualized on the basis that it would serve as an outcome measure and provide a means of describing and measuring changes in physical function for patients undergoing limb preservation surgery for sarcoma. An important component of the TESS' development was that it represent patient perceptions of their function both in content and in its completion. Irrespective of the fact that the TESS purposefully excluded certain domains of functioning, the content can be evaluated within the WHO (1980) ICIDH to determine whether the TESS is measuring the concept of impairment, disability, or handicap. This is facilitated by evaluating the questionnaire items in relation to the classification system. This availability of the classification system in the ICIDH (WHO, 1980) and the problems of the definitional boundaries of Nagi (1965, 1991) and Verbrugge (1994) were the major reasons for choosing the ICIDH (WHO, 1980) to conceptualize the TESS.

Using the impairment, disability, handicap codes (WHO, 1980), the items on the TESS measure disability (Table 1). Impairments according to the WHO classification (1980) are defined as abnormalities at the organ level and may be thought of for the
musculoskeletal system as loss of range of motion at a specific joint or loss of strength in a particular muscle group. This level of specificity in an outcome measure for the extremity sarcoma patient is undesirable. Clinical observation suggests that these patients rarely experience abnormalities at a single joint or in an isolated muscle group. For example, a patient with a soft tissue sarcoma of the thigh may experience hip and knee stiffness and weakness of muscles powering both joints. There may be oedema of the entire lower extremity. Consequently, the problem becomes one of rehabilitating the entire extremity with the goal of facilitating return of the individual to their maximal obtainable level of function. While the value of symptom control and impairment measures is recognized in monitoring clinical practice, the goal of rehabilitation is restoration of loss of functional capacity (Albrecht, 1992) and, thus, is best represented by the concept of disability. While disability follows from impairments not all impairments result in disability (WHO, 1980) and Badley (1987) has shown that there are low to moderate correlations between impairments and disabilities for individuals with stroke, multiple sclerosis, and rheumatoid and osteoarthritis. The TESS is, therefore, a measure of disability.

In summary, Nagi (1965, 1991), the WHO (1980) and Verbrugge (1994) developed disability models with the intent of providing conceptual clarity for evaluating the sequelae experienced by those with chronic disease. While none of the models provides the perfect mousetrap, they do provide a basis for consideration of the appropriate level of measurement for varying disease populations.
**Figure 1: Overview of Existing Disability Models**

<table>
<thead>
<tr>
<th>Nagi (1965, 1991)</th>
<th>Active Pathology</th>
<th>Impairment</th>
<th>Functional Limitations</th>
<th>Disability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>interruption or interference with normal processes and the effort of the organ to regain normal function</td>
<td>loss or abnormality of an anatomical, physiological or mental, or emotional nature; includes all conditions of pathology, residual losses or abnormalities after active pathology resolution, congenital abnormalities</td>
<td>result from impairments and represents an individual's loss of ability to perform the tasks and obligations of his/her usual roles and normal daily activities; refers to functioning at the organism level</td>
<td>inability or limitation in performing socially defined roles and tasks expected of the individual within a sociocultural and physical environment; includes family &amp; interpersonal relations, employment &amp; economic pursuits, education, leisure and self-care; severity is dependent on definition of situation by self and others and by the characteristics of the environment (e.g. physical and sociocultural barriers)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>World Health Organization (1980)</th>
<th>Disease</th>
<th>Impairment</th>
<th>Disability</th>
<th>Handicap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>something abnormal occurs within the individual</td>
<td>in the context of health, any loss or abnormality of psychological, physiological or anatomical structure or function; occurs at organ, tissue or structural level in body or in a functional system or body mechanism; includes functional limitations; may be temporary or permanent; exteriorization of pathological state e.g. vision, hearing loss, lack of coordination, stiffness, instability, deformity, loss of movement</td>
<td>in the context of health, any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being;</td>
<td>in the context of health, disadvantage for a given individual resulting from an impairment or a disability, that limits or prevents the fulfillment of a role that is normal (based on age, sex, social and cultural factors) for that individual; includes valuation, self and peer; represents the consequences (cultural, social, economic, environmental) of impairment and disability</td>
</tr>
</tbody>
</table>
Verbrugge (1994)  
Pathology  
biochemical and physiological abnormalities that are detected and medically labelled as disease, injury or congenital/developmental conditions; acute (< 3 mos duration) or chronic conditions; not always directly measurable; may be implied by signs symptoms  

<table>
<thead>
<tr>
<th>Impairment</th>
<th>Functional Limitations</th>
<th>Disability</th>
</tr>
</thead>
<tbody>
<tr>
<td>dysfunction and significant abnormalities in specific body systems; consequences for physical, mental, social function; may occur at primary and secondary sites; measured by patient history, clinical, laboratory, radiographic examination, home survey</td>
<td>restrictions in performing fundamental physical and mental actions used in daily life by one's age-sex group; generic actions used in many situations; tasks, include overall mobility, discrete motions and strengths, vision, hearing, communication, cognition and emotional functioning; measured by survey, observation, performance (e.g. timed) tasks</td>
<td>experienced difficulty doing activities in any domain of life (typical for age-sex group) due to a health or physical problem; roles; domains all inclusive e.g. hygiene, leisure, sleep, work, ADL's does not include dependency measures e.g. aids as they are considered &quot;interventions&quot; social context, situation specific</td>
</tr>
</tbody>
</table>

*Alignment indicates definitional overlap of categories*
Figure 2: Stylization of the Definitional Overlap by Category Heading for the Existing Disability Models

Nagi (1965, 1991)

| Active Pathology | Impairment | Functional Limitations | Disability |

World Health Organization (1980)

| Disease | Impairment | Disability | Handicap |

Verbrugge (1994)

<p>| Pathology | Impairment | Functional Limitations | Disability |</p>
<table>
<thead>
<tr>
<th>TESS Item</th>
<th>ICIDH Disability Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pants</td>
<td>Clothing</td>
</tr>
<tr>
<td>Socks</td>
<td>Other Dressing</td>
</tr>
<tr>
<td>Shoes</td>
<td>Other Dressing (hosiery)</td>
</tr>
<tr>
<td>Showering</td>
<td>Bathing Disability</td>
</tr>
<tr>
<td>Light household duties</td>
<td>Body Disposition (manual cleaning)</td>
</tr>
<tr>
<td>Gardening</td>
<td>Household (unspecified)</td>
</tr>
<tr>
<td>Meal preparation</td>
<td>Feeding</td>
</tr>
<tr>
<td>Shopping</td>
<td>Procuring Sustenance</td>
</tr>
<tr>
<td>Heavy household duties</td>
<td>Family Role</td>
</tr>
<tr>
<td>Getting in and out of a bathtub</td>
<td>Bathing (transfer)</td>
</tr>
<tr>
<td>Getting in and out of bed</td>
<td>Confining (transfer from lying)</td>
</tr>
<tr>
<td>Rising from a chair</td>
<td>Transfer from sitting</td>
</tr>
<tr>
<td>Kneeling</td>
<td>Kneeling disability</td>
</tr>
<tr>
<td>Bending</td>
<td>Body movements (retrieval)</td>
</tr>
<tr>
<td>Walking downstairs</td>
<td>Climbing stairs</td>
</tr>
<tr>
<td>Walking indoors</td>
<td>Walking</td>
</tr>
<tr>
<td>Walking outdoors</td>
<td>Traversing</td>
</tr>
<tr>
<td>Sitting</td>
<td>Other body disposition</td>
</tr>
<tr>
<td>Walking up and down hills</td>
<td>Disability in endurance</td>
</tr>
<tr>
<td>Standing</td>
<td>Other climbing</td>
</tr>
<tr>
<td>Rising from kneeling</td>
<td>Body movement (kneeling)</td>
</tr>
<tr>
<td>Getting in and out of a car</td>
<td>Personal transport</td>
</tr>
<tr>
<td>Sexual activities</td>
<td>Relations (partnership)</td>
</tr>
<tr>
<td>Work duties</td>
<td>Household family role</td>
</tr>
<tr>
<td>Working usual number of hours</td>
<td>Occupational role</td>
</tr>
<tr>
<td>Leisure activities</td>
<td>Work routine</td>
</tr>
<tr>
<td>Social activities</td>
<td>Recreation</td>
</tr>
<tr>
<td>Driving</td>
<td>Household</td>
</tr>
<tr>
<td>Walking upstairs</td>
<td>Personal transport</td>
</tr>
<tr>
<td>Sports</td>
<td>Climbing stairs</td>
</tr>
</tbody>
</table>

**TABLE 1**

Disability Codes per ICIDH for Lower Extremity TESS
APPENDIX TWO

DEVELOPMENT OF THE SHORT FORM-36 (SF-36)
DEVELOPMENT OF THE SHORT FORM-36

The Short Form 36 (SF-36) originated from the Rand Corporation Health Insurance Experiments (HIE) and the subsequent Medical Outcomes Studies (MOS). The HIE was a federally funded experimental study (1974-1982) designed to evaluate the effect of different health payment systems on the use of health services in the United States. The survey sampled approximately four thousand people, age fourteen to sixty-five. The sampling was conducted at six sites and was designed to achieve a sample representative of those living in the same geographic area (Brook, 1983). Participants were followed for three years with thirty percent followed for a further two years. Classified as "co-payers" (four payment plans with varying co-payments were collapsed) or "free" care recipients at the time of enrollment, the study found that "co-payers" made a third fewer medical visits and had a third fewer hospital admissions than those receiving "free" care. To determine if those receiving "free" care were healthier, a 108 item questionnaire assessing subjective health was administered to subjects on enrollment and again on leaving the study. Subjective general health was measured on five dimensions: physical functioning; role functioning; mental health; social contacts; and, health perceptions. This study found first that free care did not improve health status (Brook, 1983) and secondly, that self-rated health based on scales constructed from surveys had potential as reliable and valid tools for assessing changes in health status for adults and teens in the general population (Ware, 1993).

Based on the perceived success of self-rated health surveys from the HIE, the Medical Outcomes Study (MOS) incorporated a similar instrument into its methodology.
The purpose of the MOS was two-fold: 1) to determine if variations in health status were related to variations in system care, clinician specialty and clinician's technical interpersonal style; and, 2) to develop instruments for monitoring patient outcomes in medical practice, specifically a generic, self-administered questionnaire (Tarlov, 1989). It is this second purpose that is of major focus as it provides the impetus for the development of the SF-36.

The MOS included over 22,000 patients who consulted a random sample of physicians with different types of medical practices in Boston, Chicago and Los Angeles during designated screening periods between February and November 1986. Of these, 3000 patients with one or more specific medical conditions such as diabetes mellitus, hypertension, heart disease and depression, were selected for two year, longitudinal follow-up (McHorney, 1994). During this period, the number of hospitalizations and treatments were recorded and the health status of these subjects was measured repeatedly.

The outcomes in the MOS included a large set of measures, including the MOS Functioning and Well-Being Profile. This profile includes thirty-five scales with 149 items measuring physical and role functioning, social, family and sexual functioning, mobility, psychological distress/well-being, cognitive functioning, health perceptions, health distress, energy/fatigue, sleep, pain, and symptoms (Stewart, 1992). The total set of MOS measures includes 245 items.

The Short Form 20 General Health Survey (SF-20) was developed from the MOS. The intent of the SF-20 was to develop a short but comprehensive measure, covering as many dimensions of health as possible, including psychometrically valid, multi-item scales (Stewart, 1988). The HIE provided the basis for the SF-20 with eighteen items drawn from this instrument. Two single items of social functioning and bodily pain were added based on the experience of the developers (Stewart, 1988). Psychometric testing of the SF-20 was conducted and difficulties were identified that ultimately lead to the development of the SF-36. In particular, internal consistency evaluation of the multi-item
scales demonstrated an alpha of 0.76, acceptable for group evaluation. However, the single item scales of social functioning and bodily pain could not be evaluated. Ware (1987) has advocated that domains should be measured with multi-item scales. The second problem with the SF-20 is the floor effects that occur when the measure is administered to seriously ill patients (Bindman, 1990).

The goal in developing the SF-36 was to develop a generic health status measure. Generic was defined to represent the basic human values relevant to everyone's functional status and well-being; generic as in universally valued and not age, disease or treatment specific. Ware (1987) described the dimensionality of health to include physical and mental health and suggested that comprehensive measures, at minimum, include five health concepts: physical and mental health, social functioning, role functioning, and general well-being to fully incorporate the World Health Organization (WHO, 1948; WHO, 1984) definition of health in evaluating health-related quality of life. The World Health Organization (WHO, 1948; WHO, 1984) defines health as "a state of complete physical, mental and social well-being [and autonomy] and not merely the absence of disease".

The social dimension of health, however, was deliberately excluded from the SF-36 due to difficulties in measuring the construct (Ware, personal communication 1996). Attempts at measuring social health included quantification of frequency of social contacts and interactions and evaluation of quality of social contacts/supports (Ware, 1987). However, these items did not perform well in psychometric testing. The conclusion was that social well-being encompasses a much broader context than health and Bowling (1991) would concur. Bowling (1991) defines social health as encompassing well-being beyond physical and mental health, inclusive of a social support network and resources to permit coping and adaptation to life stressors.

The SF-36 represents physical and mental health dimensions (Ware, 1992) and the eight most important health concepts of the MOS are included in the measure. These eight
health concepts are representative of the minimum standards required for a comprehensive health status measurement with the addition of the concepts of bodily pain and vitality. The ninth concept is the health transition item. This conceptualization provides the underpinnings for the development of the SF-36.

Psychometric criteria of reliability and validity, specifically standards for group comparisons, were also considered in choosing the items in the SF-36. Figure 1 (adopted from Ware, 1995) diagrams the items mapped to the eight subscales and the subscales to the physical and mental health component scales. This figure provides reference for the description of the subscales and physical and mental health summary scores.

SF-36 Subscales

Physical Functioning

The complete ten item Physical Functioning scale used in the MOS was maintained in the SF-36. The full scale was considered necessary in order to represent the breadth of content inherent in physical functioning and to provide the range of minimal to severe limitations. The ten items are: vigorous activities; moderate activities; lifting or carrying groceries; climbing several flights of stairs; climbing one flight of stairs; bending/kneeling/stooping; walking more than a mile; walking several blocks; walking one block; and, bathing and dressing. Inclusion of the ten items in the SF-36 provide levels of activities between two extremes, that is basic activities of daily living to high demand or vigorous activities. The subscale was also changed to rate difficulty rather than duration (as had been done in the SF-20). This change is said to have increased the precision (Stewart, 1992; Jette, 1986) of the subscale.
Figure 1: SF-36 Measurement Model

<table>
<thead>
<tr>
<th>Items</th>
<th>Subscale</th>
<th>Summary Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a. vigorous activities</td>
<td>Physical Functioning (PF)</td>
<td>Physical Health</td>
</tr>
<tr>
<td>3b. moderate activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c. lift, carry groceries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3d. climb several flights of stairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3e. climb one flight of stairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3f. bend, kneel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3g. walk mile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3h. walk several blocks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3i. walk one block</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3j. bathe, dress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4a. cut down time</td>
<td>Role-Physical (RP)</td>
<td></td>
</tr>
<tr>
<td>4b. accomplished less</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4c. limited in kind</td>
<td>Bodily Pain (BP)</td>
<td></td>
</tr>
<tr>
<td>4d. had difficulty</td>
<td>General Health (GH)*</td>
<td></td>
</tr>
<tr>
<td>7. pain-magnitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. pain-interfere</td>
<td>Vitality (VT)*</td>
<td>Mental Health</td>
</tr>
<tr>
<td>1. general health rating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11a. sick easier</td>
<td>Social Functioning (SF)*</td>
<td></td>
</tr>
<tr>
<td>11b. as healthy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11c. health to get worse</td>
<td>Role-emotional (RE)</td>
<td></td>
</tr>
<tr>
<td>11d. health excellent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9a. pep/life</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9e. energy</td>
<td>Mental Health (MH)</td>
<td></td>
</tr>
<tr>
<td>9g. worn out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9i. tired</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. social-extent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. social-time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a. cut down time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5b. accomplished less</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5c. not careful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9b. nervous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9c. down in dumps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9d. peaceful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9f. blue/sad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9h. happy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Ware, SF-36 Physical and Mental Health Summary Scales: A User's Manual, 1994 p.3:2.

* denotes that the subscale has a factor loading greater than 0.35 with both summary measures and is significantly correlated with both the physical health and mental health component scores.
Role Functioning: Physical and Emotional

The SF-36 includes a subset of seven role functioning items from the MOS long-forms. They cover an array of role functions including limitations in, reduction of time spent and difficulty performing work and other usual activities due to physical health and due to emotional health. Specifically, the items include: cutting down the amount of time you spent on work or other activities; accomplished less than you would like; were limited in the kind of work or other activities; and, had difficulty performing the work or other activities due to physical health. In relation to mental health, the items are: cut down the amount of time you spent on work or other activities; accomplished less than you would like; and, didn’t do work or other activities as carefully as usual. These items represent more levels of role limitations than were included in the SF-20 and expands the concept of “role” for those with multiple usual roles.

Role Functioning in the SF-36 distinguishes between role limitations due to physical versus mental health problems. This latter change has been instrumental in improving the ability of the SF-36 to distinguish between groups with known medical and psychiatric problems (Hays, 1990).

Bodily Pain

The SF-36 includes two items, one asking about how much or the intensity of the pain and the second item asks how much pain interfered with normal work inside and outside the home. The SF-36 adds a second item about the intensity of bodily pain as compared to the SF-20. This item was chosen as it was the best predictor of the overall score of the Behavioral Effects of Pain Scale from the MOS (Stewart, 1992). The addition of this second item improves the reliability relative to the single item scale of the SF-20.
General Health

The five item general health scale uses the General Health Rating Index (GHRI) rather than the Current Health subscale of the SF-20. This change was made as the GHRI is more comprehensive including current health, resistance to illness and health outlook. Specifically, the five items include: a rating of general health; whether the individual thinks they get sick more easily than others; whether they are as healthy as others they know; whether the individual expects their health to get worse, better or stay the same; and, whether the individual feels their health is excellent. Ware (1992) reports that the scale is more acceptable to respondents as the items are less redundant. The questions of the GHRI are worded positively and negatively to control for response set effects. This subscale, using the GHRI items has been found to discriminate between minor and serious symptoms and to predict medical care costs and return to work following myocardial infarction (Ware, 1992).

Vitality

This four item measure was not included in the SF-20 and Ware (1992) states that it was added to "better capture differences in subjective well-being". The items were chosen from the Mental Health Inventory of the HEI and include a frequency response to: 1) did you feel full of pep? 2) did you have a lot of energy? 3) did you feel worn out? and, 4) did you feel tired?

Social Functioning

The SF-36 adds a second item to this scale over the single item of the SF-20. The two items are: the extent that physical health or emotional problems interfered with normal social activities with family, friends, neighbors, groups; and, the amount of time that physical health or emotional problems interfered with social activities (like visiting with friends, relatives). The addition of the second item improves the reliability of the subscale.
Mental Health

The five item mental health scale used in the SF-20 was maintained in the SF-36. The only modifications were in format although Ware (1992) does not explicitly describe these. The five items included are those that best predict the overall score of the thirty-eight item Mental Health Index. The items are frequency responses to: 1) have you been a very nervous person? 2) have you felt so down in the dumps that nothing could cheer you up? 3) have you felt calm and peaceful? 4) have you felt downhearted and blue? and, 5) have you been a happy person?

Reported Health Transition

The final addition to the SF-36 was a single item general health transition question asking about health changes over the past year. This item is not used in scoring any of the subscales of the SF-36.

With the exception of the health transition item, each of the subscales provides a standardized score ranging from zero to one hundred. Ware (1992) justifies the assignment of an item to a specific subscale based on the item's correlation with the overall subscale score; that is, the item has the highest correlation with the subscale of which it is a member. However, in reviewing the data presented (Ware, 1992), a number of items also have moderate correlations (range 0.40 to 0.60) with multiple subscales. Assigning items to a subscale based on the highest correlation coefficient may oversimplify the item-subscale relationship for the SF-36.

As seen in Table 1, the SF-36 correlational data from the lower extremity sarcoma sample, six months post-surgery, demonstrates a similar item to subscale pattern to that described by Ware (1992). The large number of items with moderate correlations to multiple subscales is again noted. Evaluation of the sarcoma data at multiple time points consistently produced this item-subscale relationship suggesting that this correlational pattern is not a spurious result.
TABLE 1

Item Means, Standard Deviations, and Correlations* with SF-36 Sub-Scales: Lower Extremity Six Months Post-operative Data (N=94)

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>PF</th>
<th>RP</th>
<th>BP</th>
<th>GH</th>
<th>VT</th>
<th>SF</th>
<th>RE</th>
<th>MH</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td>1.68</td>
<td>0.78</td>
<td>0.75</td>
<td>0.61</td>
<td>0.48</td>
<td>0.39</td>
<td>0.34</td>
<td>0.37</td>
<td>0.34</td>
<td>0.28</td>
</tr>
<tr>
<td>3b</td>
<td>2.26</td>
<td>0.83</td>
<td>0.88</td>
<td>0.86</td>
<td>0.57</td>
<td>0.51</td>
<td>0.53</td>
<td>0.62</td>
<td>0.57</td>
<td>0.45</td>
</tr>
<tr>
<td>3c</td>
<td>2.44</td>
<td>0.74</td>
<td>0.89</td>
<td>0.77</td>
<td>0.56</td>
<td>0.49</td>
<td>0.51</td>
<td>0.70</td>
<td>0.65</td>
<td>0.51</td>
</tr>
<tr>
<td>3d</td>
<td>2.29</td>
<td>0.76</td>
<td>0.85</td>
<td>0.74</td>
<td>0.62</td>
<td>0.50</td>
<td>0.70</td>
<td>0.63</td>
<td>0.58</td>
<td>0.51</td>
</tr>
<tr>
<td>3e</td>
<td>2.57</td>
<td>0.65</td>
<td>0.83</td>
<td>0.71</td>
<td>0.58</td>
<td>0.44</td>
<td>0.59</td>
<td>0.71</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>3f</td>
<td>2.08</td>
<td>0.73</td>
<td>0.78</td>
<td>0.60</td>
<td>0.67</td>
<td>0.45</td>
<td>0.46</td>
<td>0.56</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>3g</td>
<td>2.04</td>
<td>0.87</td>
<td>0.83</td>
<td>0.62</td>
<td>0.58</td>
<td>0.44</td>
<td>0.59</td>
<td>0.49</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>3h</td>
<td>2.32</td>
<td>0.82</td>
<td>0.89</td>
<td>0.66</td>
<td>0.56</td>
<td>0.51</td>
<td>0.58</td>
<td>0.58</td>
<td>0.55</td>
<td>0.52</td>
</tr>
<tr>
<td>3i</td>
<td>2.59</td>
<td>0.65</td>
<td>0.83</td>
<td>0.67</td>
<td>0.61</td>
<td>0.44</td>
<td>0.56</td>
<td>0.66</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>3j</td>
<td>2.83</td>
<td>0.41</td>
<td>0.60</td>
<td>0.54</td>
<td>0.23</td>
<td>0.35</td>
<td>0.42</td>
<td>0.66</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>4a</td>
<td>1.74</td>
<td>0.44</td>
<td>0.64</td>
<td>0.88</td>
<td>0.49</td>
<td>0.51</td>
<td>0.50</td>
<td>0.63</td>
<td>0.65</td>
<td>0.50</td>
</tr>
<tr>
<td>4b</td>
<td>1.66</td>
<td>0.48</td>
<td>0.77</td>
<td>0.96</td>
<td>0.59</td>
<td>0.50</td>
<td>0.53</td>
<td>0.63</td>
<td>0.73</td>
<td>0.47</td>
</tr>
<tr>
<td>4c</td>
<td>1.59</td>
<td>0.49</td>
<td>0.80</td>
<td>0.91</td>
<td>0.55</td>
<td>0.45</td>
<td>0.50</td>
<td>0.54</td>
<td>0.58</td>
<td>0.39</td>
</tr>
<tr>
<td>4d</td>
<td>1.61</td>
<td>0.49</td>
<td>0.77</td>
<td>0.92</td>
<td>0.55</td>
<td>0.51</td>
<td>0.56</td>
<td>0.58</td>
<td>0.63</td>
<td>0.42</td>
</tr>
<tr>
<td>7</td>
<td>5.10</td>
<td>1.06</td>
<td>0.64</td>
<td>0.59</td>
<td>0.94</td>
<td>0.55</td>
<td>0.54</td>
<td>0.58</td>
<td>0.63</td>
<td>0.47</td>
</tr>
<tr>
<td>8</td>
<td>4.70</td>
<td>1.18</td>
<td>0.66</td>
<td>0.57</td>
<td>0.93</td>
<td>0.58</td>
<td>0.56</td>
<td>0.53</td>
<td>0.58</td>
<td>0.43</td>
</tr>
<tr>
<td>1</td>
<td>4.17</td>
<td>0.77</td>
<td>0.59</td>
<td>0.55</td>
<td>0.60</td>
<td>0.76</td>
<td>0.52</td>
<td>0.54</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td>11a</td>
<td>4.15</td>
<td>0.98</td>
<td>0.39</td>
<td>0.38</td>
<td>0.43</td>
<td>0.61</td>
<td>0.49</td>
<td>0.46</td>
<td>0.52</td>
<td>0.45</td>
</tr>
<tr>
<td>11b</td>
<td>3.78</td>
<td>1.27</td>
<td>0.29</td>
<td>0.31</td>
<td>0.31</td>
<td>0.71</td>
<td>0.48</td>
<td>0.28</td>
<td>0.22</td>
<td>0.39</td>
</tr>
<tr>
<td>11c</td>
<td>4.24</td>
<td>0.91</td>
<td>0.24</td>
<td>0.17</td>
<td>0.21</td>
<td>0.51</td>
<td>0.22</td>
<td>0.29</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>11d</td>
<td>3.92</td>
<td>1.07</td>
<td>0.56</td>
<td>0.52</td>
<td>0.61</td>
<td>0.77</td>
<td>0.53</td>
<td>0.63</td>
<td>0.49</td>
<td>0.41</td>
</tr>
<tr>
<td>9a</td>
<td>3.95</td>
<td>1.43</td>
<td>0.53</td>
<td>0.46</td>
<td>0.46</td>
<td>0.58</td>
<td>0.74</td>
<td>0.53</td>
<td>0.46</td>
<td>0.59</td>
</tr>
<tr>
<td>9e</td>
<td>4.15</td>
<td>1.33</td>
<td>0.62</td>
<td>0.61</td>
<td>0.49</td>
<td>0.65</td>
<td>0.72</td>
<td>0.55</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>9g</td>
<td>4.70</td>
<td>1.16</td>
<td>0.53</td>
<td>0.54</td>
<td>0.52</td>
<td>0.47</td>
<td>0.81</td>
<td>0.50</td>
<td>0.49</td>
<td>0.67</td>
</tr>
<tr>
<td>9i</td>
<td>4.46</td>
<td>1.14</td>
<td>0.46</td>
<td>0.45</td>
<td>0.48</td>
<td>0.45</td>
<td>0.77</td>
<td>0.46</td>
<td>0.41</td>
<td>0.61</td>
</tr>
<tr>
<td>6</td>
<td>4.41</td>
<td>0.98</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
<td>0.53</td>
<td>0.57</td>
<td>0.79</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>4.34</td>
<td>1.10</td>
<td>0.55</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.50</td>
<td>0.84</td>
<td>0.60</td>
<td>0.50</td>
</tr>
<tr>
<td>5a</td>
<td>1.78</td>
<td>0.41</td>
<td>0.67</td>
<td>0.68</td>
<td>0.60</td>
<td>0.51</td>
<td>0.50</td>
<td>0.75</td>
<td>0.94</td>
<td>0.55</td>
</tr>
<tr>
<td>5b</td>
<td>1.75</td>
<td>0.43</td>
<td>0.64</td>
<td>0.63</td>
<td>0.61</td>
<td>0.52</td>
<td>0.51</td>
<td>0.73</td>
<td>0.95</td>
<td>0.56</td>
</tr>
<tr>
<td>5c</td>
<td>1.18</td>
<td>0.39</td>
<td>0.53</td>
<td>0.62</td>
<td>0.49</td>
<td>0.52</td>
<td>0.42</td>
<td>0.69</td>
<td>0.90</td>
<td>0.55</td>
</tr>
<tr>
<td>9b</td>
<td>5.20</td>
<td>1.04</td>
<td>0.52</td>
<td>0.46</td>
<td>0.46</td>
<td>0.48</td>
<td>0.65</td>
<td>0.54</td>
<td>0.53</td>
<td>0.84</td>
</tr>
<tr>
<td>9c</td>
<td>5.54</td>
<td>0.82</td>
<td>0.45</td>
<td>0.46</td>
<td>0.36</td>
<td>0.47</td>
<td>0.55</td>
<td>0.55</td>
<td>0.54</td>
<td>0.78</td>
</tr>
<tr>
<td>9d</td>
<td>4.54</td>
<td>1.29</td>
<td>0.37</td>
<td>0.33</td>
<td>0.38</td>
<td>0.44</td>
<td>0.60</td>
<td>0.58</td>
<td>0.40</td>
<td>0.77</td>
</tr>
<tr>
<td>9f</td>
<td>5.30</td>
<td>0.93</td>
<td>0.45</td>
<td>0.43</td>
<td>0.49</td>
<td>0.45</td>
<td>0.63</td>
<td>0.58</td>
<td>0.58</td>
<td>0.87</td>
</tr>
<tr>
<td>9h</td>
<td>4.85</td>
<td>1.03</td>
<td>0.39</td>
<td>0.29</td>
<td>0.33</td>
<td>0.52</td>
<td>0.61</td>
<td>0.49</td>
<td>0.32</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>2.71</td>
<td>0.65</td>
<td>0.09</td>
<td>0.15</td>
<td>-0.10</td>
<td>0.00</td>
<td>0.16</td>
<td>0.07</td>
<td>-0.02</td>
<td>0.12</td>
</tr>
</tbody>
</table>

*Correlations calculated using Pearson's r. Item numbers correspond to the Standard version of the SF-36. Item numbers are grouped to represent the subscales. The highest item-subscale correlation coefficient is highlighted.
The SF-36 provides a health profile. Ware (1992) emphasizes the importance of the order of presentation of the subscales: physical function; role-physical; bodily pain; general health; vitality; social function; role-emotional; and mental health. This represents a pattern from the best physical health measure to the best mental health measure. Five scales, physical function, role-physical, bodily pain, social function and role-emotional reflect absence of limitation with a perfect score of one hundred. General health, vitality and mental health, however, reflect "positive" health. Scores in the mid-range reflect no limitations whereas scores above this reflect a favourable reflection of health.

The reliability and validity of the SF-36 subscales has been reported in numerous studies and these are summarized in Ware (1993) and Jenkinson (1996). Overall, the results of these studies suggest that the SF-36 subscales have: 1) reliability sufficient for group comparisons with \( r=0.78-0.93 \) with 95% confidence intervals of the standard error of the mean of 13-32 in cross-sectional studies and test-retest reliability coefficients of 0.60 to 0.81 with 95% confidence intervals around the standard error of the mean of 19-47 (McHorney, 1995); 2) the ability to discriminate among patients with different and varying levels of medical diagnoses e.g. the mean mental health subscale score for patients with clinical depression was 46.26, sd=20.83, for chronic obstructive lung disease was 68.06, sd=19.68, for angina was 73.04, sd=18.67 and for back pain was 74.93, sd=18.62 (Ware, 1993); and, 3) that the role functioning subscales have floor effects in severely ill populations, specifically, 24% for role-physical and 18% for role-emotional (McHorney, 1994).

To follow the events in the development of the SF-36 chronologically, the International Quality of Life Assessment (IQOLA) Project needs to be introduced. The SF-36 is internationally recognized and a collaborative group is working on the translations and adaptations necessary for use of the measure in fifteen countries (Aaronson, 1992). This project has resulted in minor modifications to some of the questionnaire items to achieve a culturally appropriate measure. Notably, the project has facilitated the collection
of large data sets such that normative data for the SF-36 are now available for Britain, France, Germany, Sweden, United States and Wales.

**SF-36 Physical Component Scores and Mental Component Scores**

A more recent development in the evolution of the SF-36, is the creation of Physical and Mental Health Summary Scores (Ware, 1994). Ware (1994) suggests that there is a two factor solution for the SF-36, physical and mental health, that accounts for eighty to eighty-five percent of the variance of the eight subscales. It was hypothesized that the physical function, role-physical and bodily pain subscales would have the highest correlations with the physical health factor and that the social function, role-emotional and mental health subscales would have the highest correlations with the mental health factor. General health and vitality were anticipated to correlate with both the physical and mental health factors. Using the eigenvalue one criteria, two factors were extracted after orthogonal rotation and as hypothesized the subscales were correlated with the two factors. In the process of factor analysis, Ware (1994) has used z-transformations based on US population data such that the normalized data have a mean of 50 with a standard deviation of 10. It is to this standardized data that factor weights are applied and aggregated, producing the physical and mental health summary scores.

Ware's interpretation of these data is questionable, however, as simple structure has not been obtained through orthogonal rotation. The goal of principal component analysis is to obtain simple structure, that is the least number of factors to explain the variance with each variable (subscale in this case) correlated greater than 0.35 with only a single factor (Kleinbaum, 1988; Norman, 1993; Nunnally, 1978; Zeller, 1980). The orthogonal solution presented by Ware (1994) has three subscales, general health, vitality and social function, loading on both factors and this does not achieve simple structure or support the two factor solution for the SF-36. (This result is not surprising considering the moderate correlations of multiple items with subscale scores as noted previously.)
Even by Ware's hypothesis, these data, if one considers that there are physical and mental health concepts, is better represented by two overlapping circles. The subscales of social function, vitality and general health would be represented in the overlap indicating their correlation to both health concepts. Ware, however, justifies his interpretation based on the hypothesized construct and separates factor analysis for the purpose of determining scale structure from that of evaluating a hypothesized construct (Ware, personal communication 1996).

Other authors (Jenkinson, 1996; Wolinsky, 1996) have also factor analyzed the SF-36. Jenkinson (1996) analyzed the British data and obtained results similar to Ware (1995) with a "two factor" solution that accounted for sixty-six percent of the variability. Again the subscales of general health, vitality and social function loaded on both factors. Of note in Jenkinson's results (1996) is the eigenvalue of the physical factor (1.17). Using the eigenvalue one criteria, this barely meets the standard for maintaining factors. The physical health factor accounts for a minimal amount of the variance.

The SF-36 subscale relationships from the lower extremity sarcoma sample was also evaluated and the correlation matrix from the six month post-operative data is presented in Table 2 as an example of the results. The correlations range from 0.40 to 0.81 and suggest that all the subscales are related to each other. These results are consistent over multiple time points for the lower extremity sarcoma sample. The US population data (Ware, 1994) provided somewhat different results in that the Role-emotional and Mental Health subscales had low correlations (range 0.28 to 0.39) with the Physical Function, Role-physical and Bodily Pain subscales (these are the three subscales most strongly correlated to the Physical Health factor). This discrepancy raises the issue of the generalizability of US population data to: 1) ill subject samples; and, 2) Canadians.

The validity of the eight subscale structure was further questioned by Wolinsky (1996), who factor analyzed the 36 items of the SF-36 for a sample of the elderly, disabled with the result that the data was best represented by nine subscales. The ninth
subscales, labeled "health optimism" by Wolinsky (1996), included the items "getting sick" and "getting worse" as their own subscale. Despite Ware's interpretation of his factor analysis (Ware, 1994), the results of the above studies question the validity of the SF-36 physical and mental health summary scales and the eight subscales.

**TABLE 2**

<table>
<thead>
<tr>
<th>Correlation Matrix for the SF-36 Subscales for the Lower Extremity Sarcoma Six Month Post-operative Data (n=94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
</tr>
<tr>
<td>PF</td>
</tr>
<tr>
<td>RP</td>
</tr>
<tr>
<td>BP</td>
</tr>
<tr>
<td>GH</td>
</tr>
<tr>
<td>VT</td>
</tr>
<tr>
<td>SF</td>
</tr>
<tr>
<td>RE</td>
</tr>
<tr>
<td>MH</td>
</tr>
</tbody>
</table>

Based on their interpretation of the physical and mental health summary scales, Ware (1994) and Jenkinson (1996) describe the reliability and validity of these summary scales. However, the reader of these studies must interpret the results on their belief that the SF-36 has construct validity as a two factor model. I do not believe the physical and mental summary scales are valid. In view of the transformation undertaken to normalize the data, although there may be score differences between diagnostic groups, it is impossible to interpret what is different and how it is reflected by the items and subscales with the aggregation methods described by Ware (1994). This problem has resulted from a conceptualization of a multi-attribute measure in which the aggregation of the multiple
attributes into two component summary scores is unsupported by psychometrics, particularly externally valid criteria.

In summary, the SF-36 has evolved through much psychometric testing and is an internationally recognized generic health status measure. The eight subscales have been shown to meet the recommended criteria for group comparisons and validity has been demonstrated in many patient populations. Cautionary notes need to be made in using of the SF-36 with individuals as it does not meet the recommended standards (McHorney, 1995). The physical and mental health summary scales are not justified based on the current studies and reporting of SF-36 data should be restricted to the subscales (although Wolinsky's data (1996) suggest that there is a ninth subscale). As users of the SF-36 we need to continue to evaluate the inter-item relationships to develop a better understanding of what is being measured.
APPENDIX THREE

THE TORONTO EXTREMITY SALVAGE SCORE
TORONTO EXTREMITY SALVAGE SCORE (TESS)
GENERAL INFORMATION
(Lower Extremity)

Study Number: ____________________________

NAME:____________________________________Phone #________

Hospital Number: __ __ __ __ __ __ __ __ __

DATE OF BIRTH: _____ / _____ / _____ Age at surgery______years
      DD    MM    YY
SEX: 1___ male               2___ female

Date of Surgery: _____ / _____ / _____
      DD    MM    YY

DATE OF COMPLETION OF FORM: _____ / _____ / _____
      DD    MM    YY

1___ pre-operative
2___ 6 weeks post-operative
3___ 3 months post-operative
4___ 6 months post-operative
5___ 12 months post-operative

Site:  1___ Bone                        Side of Lesion: 1___ Right
      2___ Soft Tissue                   2___ Left

Region: 1___ proximal (i.e. proximal to elbow, proximal to knee)
        2___ distal (i.e. distal to elbow, distal to knee)

Anatomical Site: 1___ shoulder girdle/shoulder
                 2___ distal arm/elbow/proximal forearm
                 3___ distal forearm/wrist/hand
                 4___ pelvis/hip
                 5___ distal thigh/knee/proximal leg
                 6___ distal leg/ankle/foot
PATHOLOGICAL DIAGNOSIS:
1. _________ benign
2. _________ benign aggressive (GCT or fibromatosis)
3. _________ malignant
4. _________ metastatic

Histologic sub-type:__________________________________________

Treatment: 1. ___ Surgery only
2. ___ Surgery + RADS
3. ___ Surgery + Chemo.
4. ___ Surgery + RADS + Chemo.

Reconstructive procedure: 1. ___ Excision
2. ___ Excision & tissue transfer
3. ___ Curettage & bone graft
4. ___ Bulk Allograft
5. ___ Prosthesis
6. ___ Allograft prosthesis
Please answer the following questions.

1A) Please state your current work status:

1____ Employed full-time  
2____ Employed part-time  
3____ Unemployed  
4____ Retired  
5____ Student  
6____ Disabled

1B) If you are employed, please describe your current job activities (examples: desk job; truck driver):

__________________________________________________________________________

Rcode1____ active  
2____ sedentary

1C) If you are retired, unemployed, or disabled, please describe your former job activities:

__________________________________________________________________________

Rcode1____ active  
2____ sedentary

1D) If you are a student, please describe your area of study:

__________________________________________________________________________

Rcode1____ active  
2____ sedentary

1E) If you are not working do you receive financial assistance such as insurance, sick benefits or a pension?

1____ Yes  
2____ No

2) Briefly describe your leisure or recreational activities (examples: sports, gardening, reading):

__________________________________________________________________________

Rcode1____ active  
2____ sedentary

3A) Pain medication 1____ none  
2____ NSAIDS eg. anti-inflammatory drugs  
3____ non-narcotics eg. plain Tylenol  
4____ narcotics eg. percocet, morphine
3C) Frequency of pain medication:
1. not applicable i.e. no meds
2. intermittent
3. once a day
4. twice a day
5. 3 times a day
6. 4 times a day
7. more than 4 times a day

4) Describe the mobility or walking aid you use:
1. No aid
2. One cane or crutch
3. Two canes
4. Two crutches
5. Walker
6. Wheelchair
7. Motorized wheelchair or scooter

5) List the factors that limit your ability to perform your everyday activities:
1. pain
2. stiffness
3. fatigue
4. weakness
5. ROM
6. other
7. none
The following questions are about activities commonly performed in daily life. Each question has two parts.

A) The first part asks that you mark each item (as in the examples below) opposite the description that best describes your ability to perform each task during the past week. Some activities will be extremely easy for you to do, others will be extremely difficult or impossible.

EXAMPLE
Riding a bicycle is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

You should choose the response "impossible to do..." if the activity is something that you normally do in your daily activities but are now unable to do because of physical limitations such as weakness, stiffness or pain.

B) The second part of the question asks that you rate the activity from 1 to 4, based on the degree of importance you place on being able to perform the activity without any difficulty. For example, if it is extremely important that you perform the activity, you might rate the importance as "4, extremely important"; if the activity is unimportant to you, you might rate it as "1, totally unimportant".

EXAMPLE
Being able to ride a bicycle without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
If you do not perform an activity as part of your normal lifestyle you would choose the response "99" to indicate that the item is not applicable. If you choose response "99" you do not need to complete part B of the question.

Mark all items ensuring that you choose the description that most accurately describes your abilities in the past week.
The following questions ask about your ability to perform activities that are common to every day life. Considering the amount of difficulty you have performing the activity due to the current problem you are having with your leg, please answer the questions by choosing the answer that best describes your ability to do the activity over the past week.

1A) Putting on a pair of pants is:
   1_____impossible to do.
   2_____extremely difficult.
   3_____moderately difficult.
   4_____a little bit difficult.
   5_____not at all difficult.

   99_____This task is not applicable for me.

1B) Being able to put on a pair of pants without difficulty is:
   1_____totally unimportant.
   2_____somewhat important.
   3_____very important.
   4_____extremely important.

2A) Putting on shoes is:
   1_____impossible to do.
   2_____extremely difficult.
   3_____moderately difficult.
   4_____a little bit difficult.
   5_____not at all difficult.

   99_____This task is not applicable for me.

2B) Being able to put on shoes without difficulty is:
   1_____totally unimportant.
   2_____somewhat important.
   3_____very important.
   4_____extremely important.
3A) Putting on socks or stockings is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

3B) Being able to put on socks or stockings without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

4A) Showering is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

4B) Showering without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
5A) Light household chores such as tidying and dusting are:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

5B) Being able to do light household chores without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

6A) Gardening and yard work are:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

6B) Being able to do gardening and yard work without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
7A) Preparing and serving meals is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

7B) Being able to prepare and serve meals without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

8A) Going shopping is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

8B) Being able to go shopping without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
9A) Heavy household chores such as vacuuming and moving furniture is:

1_____impossible to do.
2_____extremely difficult.
3_____moderately difficult.
4_____a little bit difficult.
5_____not at all difficult.

99____This task is not applicable for me.

9B) Being able to do heavy household chores without difficulty is:

1_____totally unimportant.
2_____somewhat important.
3_____very important.
4_____extremely important.

10A) Getting in and out of the bath tub is:

1_____impossible to do.
2_____extremely difficult.
3_____moderately difficult.
4_____a little bit difficult.
5_____not at all difficult.

99____This task is not applicable for me.

10B) Being able to get in and out of the bath tub without difficulty is:

1_____totally unimportant.
2_____somewhat important.
3_____very important.
4_____extremely important.
11A) Getting out of bed is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

11B) Being able to get out of bed without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

12A) Rising from a chair is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

12B) Being able to rise from a chair without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
13A) Kneeling is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

13B) Being able to kneel without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

14A) Bending to pick something up off the floor is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

14B) Being able to bend to pick something up off the floor without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
15A) Walking upstairs is:

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

15B) Being able to walk upstairs without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.

16A) Walking downstairs is:

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

16B) Being able to walk down stairs without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.
17A) Driving is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

17B) Being able to drive without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

18A) Walking within the house is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

18B) Being able to walk within the house without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
19A) Walking outdoors is:

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

19B) Being able to walk outdoors without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.

20A) Sitting is:

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

20B) Being able to sit without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.
21A) Walking up or down hills or a ramp is:

1____impossible to do.
2____extremely difficult.
3____moderately difficult.
4____a little bit difficult.
5____not at all difficult.

99____This task is not applicable for me.

21B) Being able to walk up or down hills or a ramp without difficulty is:

1____totally unimportant.
2____somewhat important.
3____very important.
4____extremely important.

22A) Standing upright is:

1____impossible to do.
2____extremely difficult.
3____moderately difficult.
4____a little bit difficult.
5____not at all difficult.

99____This task is not applicable for me.

22B) Being able to stand upright without difficulty is:

1____totally unimportant.
2____somewhat important.
3____very important.
4____extremely important.
23A) Getting up from kneeling is:

1____impossible to do.
2____extremely difficult.
3____moderately difficult.
4____a little bit difficult.
5____not at all difficult.

99____This task is not applicable for me.

23B) Being able to get up from kneeling without difficulty is:

1____totally unimportant.
2____somewhat important.
3____very important.
4____extremely important.

24A) Getting in and out of a car is:

1____impossible to do.
2____extremely difficult.
3____moderately difficult.
4____a little bit difficult.
5____not at all difficult.

99____This task is not applicable for me.

24B) Being able to get in and out of a car without difficulty is:

1____totally unimportant.
2____somewhat important.
3____very important.
4____extremely important.
25A) Participating in sexual activities is:

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

25B) Being able to participate in sexual activities without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.

26A) Completing my usual duties at work is: (Work includes both a job outside the home and as a homemaker.)

1 ___ impossible to do.
2 ___ extremely difficult.
3 ___ moderately difficult.
4 ___ a little bit difficult.
5 ___ not at all difficult.

99 ___ This task is not applicable for me.

26B) Being able to complete my usual duties at work or as a homemaker without difficulty is:

1 ___ totally unimportant.
2 ___ somewhat important.
3 ___ very important.
4 ___ extremely important.
27A) Working my usual number of hours is: (Working includes both a job outside the home and as a homemaker.)

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

27B) Being able to work my usual number of hours without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

28A) Participating in my usual leisure activities is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

28B) Being able to participate in my usual leisure activities without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
29A) Socializing with friends and family is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

29B) Being able to socialize with friends and family without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

30A) Participating in my usual sporting activities is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

99. This task is not applicable for me.

30B) Being able to participate in my usual sporting activities without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
Are there any activities that you find difficult and are important to you that have not been included in this questionnaire? If so please describe the activity and the amount of difficulty you have had during the past week doing the activity due to the current problem in your leg. Rate how important it is to you that you be able to perform this activity without difficulty.

31A) Activity:__________________________________________

This activity is:  1____ impossible to do.
                    2____ extremely difficult.
                    3____ moderately difficult.
                    4____ a little bit difficult.
                    5____ not at all difficult.

31B) Being able to do this activity without difficulty is:

                    1____ totally unimportant.
                    2____ somewhat important.
                    3____ very important.
                    4____ extremely important.

32A) Activity:__________________________________________

This activity is:  1____ impossible to do.
                    2____ extremely difficult.
                    3____ moderately difficult.
                    4____ a little bit difficult.
                    5____ not at all difficult.

32B) Being able to do this activity without difficulty is:

                    1____ totally unimportant.
                    2____ somewhat important.
                    3____ very important.
                    4____ extremely important.
33A) Activity: _________________________________________

This activity is:
1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

33B) Being able to do this activity without difficulty is:
1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.

34A) Activity: _________________________________________

This activity is:
1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

34B) Being able to do this activity without difficulty is:
1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
35A) Activity: __________________________

This activity is:

1. impossible to do.
2. extremely difficult.
3. moderately difficult.
4. a little bit difficult.
5. not at all difficult.

35B) Being able to do this activity without difficulty is:

1. totally unimportant.
2. somewhat important.
3. very important.
4. extremely important.
1) Considering all the activities in which I participate in daily life, I would rate the ability to perform these activities during the past week as:

1____ impossible to do.
2____ extremely difficult.
3____ moderately difficult.
4____ a little bit difficult.
5____ not at all difficult.

2) I would rate myself as being:

1____ completely disabled
2____ severely disabled.
3____ moderately disabled.
4____ mildly disabled.
5____ not at all disabled.

Please comment below on any activities you find difficult to perform or on any other difficulties you experience due to the problem you currently have in your leg that you feel are important and have not been asked about in this questionnaire.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Please check to make sure that you have answered all the questions.

Thank you for taking the time to answer these questions.
APPENDIX FOUR

FACTOR ANALYSIS RESULTS OVER TIME
LOWER EXTREMITY SIX MONTHS POST-OPERATIVELY:
SUMMARY OF THE CORRELATION MATRIX

<table>
<thead>
<tr>
<th>ITEM</th>
<th>MEAN (sd) CORRELATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>standing</td>
<td>.23(.21)</td>
</tr>
<tr>
<td>driving</td>
<td>.24(.18)</td>
</tr>
<tr>
<td>sexual activities</td>
<td>.27(.17)</td>
</tr>
<tr>
<td>work duties</td>
<td>.33(.18)</td>
</tr>
<tr>
<td>sitting</td>
<td>.37(.17)</td>
</tr>
<tr>
<td>light household</td>
<td>.38(.18)</td>
</tr>
<tr>
<td>meal preparation</td>
<td>.38(.17)</td>
</tr>
<tr>
<td>gardening/yard work</td>
<td>.39(.15)</td>
</tr>
<tr>
<td>work hours</td>
<td>.39(.16)</td>
</tr>
<tr>
<td>social activities</td>
<td>.40(.16)</td>
</tr>
<tr>
<td>pants</td>
<td>.41(.21)</td>
</tr>
<tr>
<td>walking indoors</td>
<td>.42(.18)</td>
</tr>
<tr>
<td>showering</td>
<td>.42(.16)</td>
</tr>
<tr>
<td>socks</td>
<td>.43(.18)</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>.43(.19)</td>
</tr>
<tr>
<td>shoes</td>
<td>.43(.20)</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>.44(.15)</td>
</tr>
<tr>
<td>getting in/out of a car</td>
<td>.45(.15)</td>
</tr>
<tr>
<td>sports</td>
<td>.45(.15)</td>
</tr>
<tr>
<td>leisure activities</td>
<td>.47(.17)</td>
</tr>
<tr>
<td>bending</td>
<td>.48(.18)</td>
</tr>
<tr>
<td>kneeling</td>
<td>.48(.14)</td>
</tr>
<tr>
<td>bathtub</td>
<td>.49(.17)</td>
</tr>
<tr>
<td>hills</td>
<td>.50(.16)</td>
</tr>
<tr>
<td>going downstairs</td>
<td>.51(.19)</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>.51(.18)</td>
</tr>
<tr>
<td>heavy household</td>
<td>.52(.17)</td>
</tr>
<tr>
<td>shopping</td>
<td>.52(.17)</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>.53(.18)</td>
</tr>
<tr>
<td>going upstairs</td>
<td>.53(.18)</td>
</tr>
</tbody>
</table>

Mean correlation of the matrix is 0.41, sd=0.15.
**PRINCIPAL COMPONENT ANALYSIS ORTHOGONAL ROTATION**

**SIX MONTHS POST-OPERATIVELY**

*Bolded items load on more than one factor*

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>pants</td>
<td>.31</td>
<td>-.01</td>
<td>.10</td>
<td>-.08</td>
<td>.77</td>
<td>.18</td>
<td>.74</td>
</tr>
<tr>
<td>shoes</td>
<td>.25</td>
<td>.03</td>
<td>.14</td>
<td>.05</td>
<td>.83</td>
<td>.12</td>
<td>.79</td>
</tr>
<tr>
<td>socks</td>
<td>.28</td>
<td>.20</td>
<td>.09</td>
<td>.00</td>
<td>.73</td>
<td>.10</td>
<td>.67</td>
</tr>
<tr>
<td>shower</td>
<td>.40</td>
<td>.04</td>
<td>.16</td>
<td>.13</td>
<td>.22</td>
<td>.54</td>
<td>.55</td>
</tr>
<tr>
<td>*light household</td>
<td>.30</td>
<td>.61</td>
<td>-.02</td>
<td>-.15</td>
<td>.17</td>
<td>.37</td>
<td>.65</td>
</tr>
<tr>
<td>garden/yard</td>
<td>.44</td>
<td>.52</td>
<td>.17</td>
<td>.27</td>
<td>.17</td>
<td>-.07</td>
<td>.60</td>
</tr>
<tr>
<td>meal preparation</td>
<td>.14</td>
<td>.55</td>
<td>.00</td>
<td>-.10</td>
<td>.18</td>
<td>.54</td>
<td>.66</td>
</tr>
<tr>
<td>shopping</td>
<td>.47</td>
<td>.26</td>
<td>.03</td>
<td>.28</td>
<td>.43</td>
<td>.42</td>
<td>.73</td>
</tr>
<tr>
<td>heavy household</td>
<td>.58</td>
<td>.40</td>
<td>.10</td>
<td>.10</td>
<td>.35</td>
<td>.29</td>
<td>.82</td>
</tr>
<tr>
<td>bathtub</td>
<td>.28</td>
<td>.23</td>
<td>.11</td>
<td>.10</td>
<td>.61</td>
<td>.35</td>
<td>.66</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>.08</td>
<td>.11</td>
<td>.28</td>
<td>.24</td>
<td>.39</td>
<td>.61</td>
<td>.68</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>.26</td>
<td>-.06</td>
<td>.25</td>
<td>.03</td>
<td>.55</td>
<td>.42</td>
<td>.62</td>
</tr>
<tr>
<td>kneeling</td>
<td>.76</td>
<td>.21</td>
<td>.04</td>
<td>.13</td>
<td>.28</td>
<td>.20</td>
<td>.76</td>
</tr>
<tr>
<td>bending</td>
<td>.50</td>
<td>.26</td>
<td>.55</td>
<td>.01</td>
<td>.18</td>
<td>.21</td>
<td>.70</td>
</tr>
<tr>
<td>upstairs</td>
<td>.51</td>
<td>.20</td>
<td>.14</td>
<td>.02</td>
<td>.46</td>
<td>.46</td>
<td>.75</td>
</tr>
<tr>
<td>downstairs</td>
<td>.60</td>
<td>.14</td>
<td>.14</td>
<td>.14</td>
<td>.35</td>
<td>.40</td>
<td>.71</td>
</tr>
<tr>
<td>driving</td>
<td>.11</td>
<td>.06</td>
<td>.17</td>
<td>.87</td>
<td>.00</td>
<td>.10</td>
<td>.81</td>
</tr>
<tr>
<td>walking indoors</td>
<td>.34</td>
<td>.02</td>
<td>.24</td>
<td>.02</td>
<td>.12</td>
<td>.74</td>
<td>.74</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>.42</td>
<td>.13</td>
<td>.21</td>
<td>.13</td>
<td>.36</td>
<td>.64</td>
<td>.80</td>
</tr>
<tr>
<td>sitting</td>
<td>-.05</td>
<td>.12</td>
<td>.75</td>
<td>.32</td>
<td>.37</td>
<td>.10</td>
<td>.84</td>
</tr>
<tr>
<td>hills</td>
<td>.52</td>
<td>.15</td>
<td>.57</td>
<td>-.05</td>
<td>.31</td>
<td>.26</td>
<td>.78</td>
</tr>
<tr>
<td>standing</td>
<td>.05</td>
<td>.06</td>
<td>.89</td>
<td>.04</td>
<td>-.13</td>
<td>.11</td>
<td>.83</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>.75</td>
<td>.19</td>
<td>.07</td>
<td>.16</td>
<td>.30</td>
<td>.28</td>
<td>.79</td>
</tr>
<tr>
<td>getting in/out of car</td>
<td>.26</td>
<td>.14</td>
<td>.69</td>
<td>.15</td>
<td>.35</td>
<td>.13</td>
<td>.73</td>
</tr>
<tr>
<td>sex</td>
<td>.21</td>
<td>.15</td>
<td>.09</td>
<td>.78</td>
<td>.05</td>
<td>.04</td>
<td>.69</td>
</tr>
<tr>
<td>work duties</td>
<td>.21</td>
<td>.79</td>
<td>.21</td>
<td>.16</td>
<td>.02</td>
<td>-.01</td>
<td>.74</td>
</tr>
<tr>
<td>work hours</td>
<td>.31</td>
<td>.67</td>
<td>.23</td>
<td>.25</td>
<td>.07</td>
<td>.05</td>
<td>.67</td>
</tr>
<tr>
<td>leisure activities</td>
<td>.37</td>
<td>.35</td>
<td>.16</td>
<td>-.03</td>
<td>.50</td>
<td>.27</td>
<td>.60</td>
</tr>
<tr>
<td>social activities</td>
<td>.06</td>
<td>.46</td>
<td>.05</td>
<td>.27</td>
<td>.54</td>
<td>.36</td>
<td>.70</td>
</tr>
<tr>
<td>sports</td>
<td>.69</td>
<td>.25</td>
<td>.23</td>
<td>.21</td>
<td>.15</td>
<td>.09</td>
<td>.67</td>
</tr>
<tr>
<td><strong>Eigenvalue</strong></td>
<td><strong>4.92</strong></td>
<td><strong>2.95</strong></td>
<td><strong>3.30</strong></td>
<td><strong>2.08</strong></td>
<td><strong>4.05</strong></td>
<td><strong>4.02</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td><strong>16.39</strong></td>
<td><strong>9.85</strong></td>
<td><strong>10.99</strong></td>
<td><strong>6.95</strong></td>
<td><strong>13.50</strong></td>
<td><strong>13.41</strong></td>
<td></td>
</tr>
</tbody>
</table>
LOWER EXTREMITY TWELVE MONTHS POST-OPERATIVELY:
SUMMARY OF THE CORRELATION MATRIX

<table>
<thead>
<tr>
<th>ITEM</th>
<th>MEAN (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>standing</td>
<td>.37 (.18)</td>
</tr>
<tr>
<td>driving</td>
<td>.21 (.20)</td>
</tr>
<tr>
<td>sexual activities</td>
<td>.25 (.17)</td>
</tr>
<tr>
<td>work duties</td>
<td>.41 (.14)</td>
</tr>
<tr>
<td>sitting</td>
<td>.39 (.19)</td>
</tr>
<tr>
<td>light household</td>
<td>.30 (.21)</td>
</tr>
<tr>
<td>meal preparation</td>
<td>.35 (.17)</td>
</tr>
<tr>
<td>gardening/yard work</td>
<td>.35 (.15)</td>
</tr>
<tr>
<td>work hours</td>
<td>.38 (.17)</td>
</tr>
<tr>
<td>social activities</td>
<td>.40 (.16)</td>
</tr>
<tr>
<td>pants</td>
<td>.44 (.17)</td>
</tr>
<tr>
<td>walking indoors</td>
<td>.38 (.18)</td>
</tr>
<tr>
<td>showering</td>
<td>.37 (.19)</td>
</tr>
<tr>
<td>socks</td>
<td>.51 (.17)</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>.44 (.16)</td>
</tr>
<tr>
<td>shoes</td>
<td>.50 (.16)</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>.40 (.17)</td>
</tr>
<tr>
<td>getting in/out of a car</td>
<td>.52 (.17)</td>
</tr>
<tr>
<td>sports</td>
<td>.40 (.16)</td>
</tr>
<tr>
<td>leisure activities</td>
<td>.39 (.16)</td>
</tr>
<tr>
<td>bending</td>
<td>.49 (.16)</td>
</tr>
<tr>
<td>kneeling</td>
<td>.47 (.17)</td>
</tr>
<tr>
<td>bathtub</td>
<td>.40 (.16)</td>
</tr>
<tr>
<td>hills</td>
<td>.52 (.19)</td>
</tr>
<tr>
<td>going downstairs</td>
<td>.48 (.18)</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>.51 (.18)</td>
</tr>
<tr>
<td>heavy household</td>
<td>.47 (.16)</td>
</tr>
<tr>
<td>shopping</td>
<td>.44 (.17)</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>.51 (.16)</td>
</tr>
<tr>
<td>going upstairs</td>
<td>.53 (.18)</td>
</tr>
</tbody>
</table>

Mean correlation of the matrix is 0.42, sd=0.08.
### PRINCIPAL COMPONENT ANALYSIS ORTHOGONAL ROTATION

**TWELVE MONTHS POST-OPERATIVELY**

*Bolded items load on more than one factor*

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>pants</td>
<td>.69</td>
<td>.25</td>
<td>-.22</td>
<td>.29</td>
<td>-.06</td>
<td>-.03</td>
<td>-.27</td>
<td>0.75</td>
</tr>
<tr>
<td>shoes</td>
<td>.79</td>
<td>-.18</td>
<td>-.05</td>
<td>.09</td>
<td>-.22</td>
<td>.00</td>
<td>-.24</td>
<td>0.75</td>
</tr>
<tr>
<td>socks</td>
<td>.80</td>
<td>-.17</td>
<td>-.05</td>
<td>.09</td>
<td>-.21</td>
<td>-.03</td>
<td>-.01</td>
<td>0.76</td>
</tr>
<tr>
<td>shower</td>
<td>.58</td>
<td>.04</td>
<td>.31</td>
<td>-.31</td>
<td>-.44</td>
<td>.00</td>
<td>-.17</td>
<td>0.76</td>
</tr>
<tr>
<td>*light household</td>
<td>.44</td>
<td>.59</td>
<td>-.52</td>
<td>.10</td>
<td>-.08</td>
<td>.01</td>
<td>.08</td>
<td>0.85</td>
</tr>
<tr>
<td>garden/yard</td>
<td>.53</td>
<td>.33</td>
<td>.10</td>
<td>.17</td>
<td>.19</td>
<td>-.16</td>
<td>.21</td>
<td>0.54</td>
</tr>
<tr>
<td>meal preparation</td>
<td>.53</td>
<td>.41</td>
<td>-.37</td>
<td>-.01</td>
<td>.17</td>
<td>-.10</td>
<td>.26</td>
<td>0.69</td>
</tr>
<tr>
<td>shopping</td>
<td>.70</td>
<td>.06</td>
<td>-.10</td>
<td>-.23</td>
<td>.14</td>
<td>.01</td>
<td>-.01</td>
<td>0.57</td>
</tr>
<tr>
<td>heavy household</td>
<td>.72</td>
<td>.46</td>
<td>-.21</td>
<td>.02</td>
<td>.10</td>
<td>-.26</td>
<td>.00</td>
<td>0.85</td>
</tr>
<tr>
<td>bathtub</td>
<td>.62</td>
<td>-.28</td>
<td>.01</td>
<td>-.05</td>
<td>.19</td>
<td>.11</td>
<td>.14</td>
<td>0.61</td>
</tr>
<tr>
<td>getting in/out of bed</td>
<td>.62</td>
<td>-.28</td>
<td>.01</td>
<td>-.04</td>
<td>.19</td>
<td>.11</td>
<td>.30</td>
<td>0.51</td>
</tr>
<tr>
<td>rising from a chair</td>
<td>.68</td>
<td>-.14</td>
<td>-.11</td>
<td>.22</td>
<td>-.27</td>
<td>-.15</td>
<td>.12</td>
<td>0.66</td>
</tr>
<tr>
<td>kneeling</td>
<td>.71</td>
<td>.22</td>
<td>.38</td>
<td>.04</td>
<td>.10</td>
<td>-.26</td>
<td>-.14</td>
<td>0.80</td>
</tr>
<tr>
<td>bending</td>
<td>.77</td>
<td>.09</td>
<td>-.06</td>
<td>.00</td>
<td>-.17</td>
<td>-.20</td>
<td>-.19</td>
<td>0.72</td>
</tr>
<tr>
<td>upstairs</td>
<td>.84</td>
<td>-.10</td>
<td>.09</td>
<td>-.30</td>
<td>-.02</td>
<td>.00</td>
<td>.12</td>
<td>0.84</td>
</tr>
<tr>
<td>downstairs</td>
<td>.76</td>
<td>-.09</td>
<td>.23</td>
<td>-.30</td>
<td>-.23</td>
<td>-.05</td>
<td>.16</td>
<td>0.81</td>
</tr>
<tr>
<td>driving</td>
<td>.29</td>
<td>.01</td>
<td>.66</td>
<td>.52</td>
<td>-.06</td>
<td>.18</td>
<td>.15</td>
<td>0.86</td>
</tr>
<tr>
<td>walking indoors</td>
<td>.59</td>
<td>-.40</td>
<td>-.04</td>
<td>-.01</td>
<td>.07</td>
<td>.02</td>
<td>.11</td>
<td>0.53</td>
</tr>
<tr>
<td>walking outdoors</td>
<td>.80</td>
<td>-.19</td>
<td>.02</td>
<td>-.11</td>
<td>.11</td>
<td>.06</td>
<td>.09</td>
<td>0.70</td>
</tr>
<tr>
<td>sitting</td>
<td>.60</td>
<td>-.45</td>
<td>-.19</td>
<td>.40</td>
<td>.10</td>
<td>.05</td>
<td>.00</td>
<td>0.78</td>
</tr>
<tr>
<td>hills</td>
<td>.83</td>
<td>-.18</td>
<td>.05</td>
<td>.28</td>
<td>.10</td>
<td>-.10</td>
<td>-.02</td>
<td>0.92</td>
</tr>
<tr>
<td>standing</td>
<td>.58</td>
<td>-.32</td>
<td>.00</td>
<td>-.12</td>
<td>.29</td>
<td>.28</td>
<td>.34</td>
<td>0.74</td>
</tr>
<tr>
<td>up from kneeling</td>
<td>.73</td>
<td>.11</td>
<td>.43</td>
<td>-.02</td>
<td>.16</td>
<td>-.25</td>
<td>-.07</td>
<td>0.92</td>
</tr>
<tr>
<td>getting in/out of car</td>
<td>.82</td>
<td>-.16</td>
<td>-.18</td>
<td>.09</td>
<td>.00</td>
<td>-.14</td>
<td>.10</td>
<td>0.76</td>
</tr>
<tr>
<td>sex</td>
<td>.35</td>
<td>.29</td>
<td>.25</td>
<td>.67</td>
<td>.00</td>
<td>.16</td>
<td>.03</td>
<td>0.74</td>
</tr>
<tr>
<td>work duties</td>
<td>.62</td>
<td>.35</td>
<td>.01</td>
<td>.02</td>
<td>-.17</td>
<td>-.52</td>
<td>-.00</td>
<td>0.81</td>
</tr>
<tr>
<td>work hours</td>
<td>.57</td>
<td>.30</td>
<td>-.09</td>
<td>-.27</td>
<td>-.09</td>
<td>.61</td>
<td>-.13</td>
<td>0.90</td>
</tr>
<tr>
<td>leisure activities</td>
<td>.60</td>
<td>.07</td>
<td>.05</td>
<td>-.10</td>
<td>.42</td>
<td>.11</td>
<td>-.42</td>
<td>0.74</td>
</tr>
<tr>
<td>social activities</td>
<td>.62</td>
<td>-.19</td>
<td>-.19</td>
<td>.24</td>
<td>.32</td>
<td>.07</td>
<td>-.39</td>
<td>0.77</td>
</tr>
<tr>
<td>sports</td>
<td>.61</td>
<td>.39</td>
<td>.29</td>
<td>-.06</td>
<td>.18</td>
<td>.02</td>
<td>.07</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Eigenvalue</strong></td>
<td>4.76</td>
<td>3.17</td>
<td>4.40</td>
<td>1.97</td>
<td>3.81</td>
<td>2.05</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>15.87</td>
<td>10.59</td>
<td>14.66</td>
<td>6.58</td>
<td>12.70</td>
<td>6.83</td>
<td>6.76</td>
<td></td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY

Aaronson NK, Bullinger M, and Ahmedzai S. A Modular Approach to Quality-of-Life Assessment in Cancer Clinical Trials. Recent Results in Cancer Research, 1988; 111: 231-249.


Davis AM, Wright JG, Williams JI, Bombardier C, Griffin A, and Bell RS. Development of a Measure of Physical Function for Patients with Bone and Soft Tissue Sarcomas. Quality of Life Research, 1996; 5(5): 508-516.


Davis AM, Young NL, Williams JI, Bell RS, Bombardier C, and Wright JG. Clinimetrics and Psychometrics: Complimentary Approaches to Questionnaire Development. Journal of Clinical Epidemiology, submitted 1996.


Ganz PA, Schag, CAC, Lee JJ and Sim MS. The CARES: A Generic Measure of Health-related Quality of Life for Patients with Cancer. Quality of Life Research, 1992; 1: 19-29.


Gross AE. Use of the Toronto Extremity Salvage Score: Lower Extremity in Hip Revision. Personal communication.


Jagial S. Use of the Toronto Extremity Salvage Score in Hip Fracture. Personal communication, 1996.


McHorney CA, Ware JE Jr. and Raczek AE. The MOS SF-36 Short-Form Health Survey (SF-36): II. Psychometric and Clinical Test of Validity in Measuring Physical and Mental Health Constructs. Medical Care, 1993; 31 (3): 247-263.


Norman GR and Streiner DL. *Biostatistics The Bare Essentials*. Toronto: Mosby, 1993.


Osoba D. *Effect of Cancer on Quality of Life*. Boca Raton, FL: CRC Press, 1991


Ware JE Jr. The Association of Biostatistical Analysts Conference. Toronto, personal communication, October 12, 1996.


Wolinsky FD and Stump TE. A Measurement Model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a Clinical Sample of Disadvantaged, Older, Black and White Men and Women. Medical Care, 1996; 34(6): 537-548.


