INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

THE UNIVERSITY OF ALBERTA

INTER-RATER RELIABILITY OF FORMALLY TRAINED AND SELF-TRAINED RATERS USING THE EDMONTON FUNCTIONAL ASSESSMENT TOOL

BY

TERRY RAE KAASA

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF PHYSICAL THERAPY

EDMONTON, ALBERTA

FALL 1998

UNIVERSITY OF ALBERTA

LIBRARY RELEASE FORM

Name of Author: Terry Rae Kaasa

Title of Thesis: Inter-rater Reliability of Formally Trained and Self-trained Raters using the Edmonton Functional Assessment Tool

Degree: Master of Science

Year this Degree Granted: 1998

7214-119 Street,
Edmonton, Alberta
Canada, T6G 1V6

Date: October 2, 1998

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled INTER-RATER RELIABILITY OF FORMALLY TRAINED AND SELF-TRAINED RATERS USING THE EDMONTON FUNCTIONAL ASSESSMENT TOOL submitted by TERRY RAE KAASA in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE.

Jean Wessel
Dr. Jean Wessel (Supervisor)

Johanna Darrah
Dr. Johanna Darrah

Dr. Eduardo Bruera

Date: Oct 1, 1998

ABSTRACT

The primary objective of this study was to determine the inter-rater reliability of the revised version of the Edmonton Functional Assessment Tool (EFAT-2). A second objective was to determine whether both formally trained and self-trained therapists had an acceptable level of inter-rater reliability. The EFAT-2 was administered to consenting palliative care (PC) patients by one of two independent physical therapist rater pairs; one pair self-trained (R1, R2) and the other formally trained (R3, R4). The intraclass correlation [ICC(1,1)] for R1, R2 was 0.97 (95% CI 0.97,0.99) and for R3, R4 was 0.95 (95% CI 0.95, 0.99). The standard error of measurement (SEM) was 1.09 and 1.44 respectively. The Kappa statistic for the rater pairs on individual EFAT items ranged from 0.51 – 0.95. The results suggest that both formally trained and self-trained therapists using the EFAT-2 obtain an acceptable level of inter-rater reliability.

Intention
Live with intention.
Walk to the edge.
Continue to learn..... .....

To my family, and to Joan Loomis M. Ed.,
who have always had faith in my ability and who have
always encouraged and supported me in my educational pursuits.

Acknowledgments

The past two years have been an exciting and challenging time for me. My experience was enhanced by the many individuals who shared freely of their time and expertise:

The research participants, who gave of their time.

My supervisor, Dr. Jean Wessel, whose insightful thoughts and words of encouragement patiently guided me through this experience. Thank you for your assistance and belief in me.

My committee member, Dr. Johanna Darrah, whose thoughtful reviews offered new perspectives and challenged me to think broadly.

My committee member, Dr. Eduardo Bruera, whose vision and passion for research encouraged my learning pursuits.

The organization that supported my research, the Caritas Health Group, with thanks to the Research Steering Committee, Dr. G.F. Mac Donald, chair.

The physical therapists who collected the data, B. O'Brien, K. Elliott, E. Ferguson, and S. Harke, for their participation, support, and expertise.

And to my family, Rod, Kristin, Ben and Adam who shared this experience with me, constantly encouraging and supporting me.

Thank you.

TABLE OF CONTENTS

Page

LIST OF TABLES

ABBREVIATIONS

ADL	Activities of Daily Living
ANOVA	Analysis of Variance
ECOG	Eastern Cooperative Oncology Group Scale
EFAT	Edmonton Functional Assessment Tool
EGH	Edmonton General Hospital
ERSS	Edinburgh Rehabilitation Status Scale
FIM	Functional Independence Measure
GNCHC	Grey Nun's Community Hospital and Health Center
ICC	Intraclass Correlation
KPS	Karnofsky Performance Status Scale
PC	Palliative Care
PCU	Palliative Care Unit
PT	Physical Therapy/Therapist
PS	Performance Status
MMSE	Mini Mental Status Exam
SEM	Standard Error of Measurement
X^2	Chi Square

CHAPTER ONE

INTRODUCTION

1.1 Problem Statement

Cancer, its treatment and palliative management has a profound effect on a person's ability to perform activities essential to life.[1] The majority of patients with terminal cancer have an impaired ability to perform everyday functions during various stages of the disease. Rehabilitation offers interventions to achieve an optimal level of functioning within the limits imposed by the disease process.[2] Rehabilitation offers patients the opportunity to find ways "to live until death".[2]

Though rehabilitation therapists observe changes in the way patients function during palliative care (PC), such changes are not well described in the literature.[3] Literature reviews of performance status (PS) instruments used in oncology and functional assessment scales in rehabilitation fail to reveal an instrument suitable for measuring functional status in the palliative care population. A comprehensive evaluation of functional status is needed to describe how patients with terminal cancer perform activities important to everyday functioning at various intervals during their care. An evaluation is also needed to measure how rehabilitation influences change in the way a patient functions in PC. An instrument that accurately measures functional status will allow clinicians to assess their patients comprehensively. Therefore development of an instrument to meet the needs of patients in PC was undertaken.

The purpose of the Edmonton Functional Assessment Tool (EFAT) is to evaluate the extent to which patients perform functions that are fundamental to interacting with people and moving in the PC environment, even when severely limited in the terminal

phase of cancer. The EFAT was designed to describe the functional status of patients throughout a declining course in PC for both rehabilitation planning and evaluation.[4] Accurate identification of specific problems that patients experience in a PC setting will drive rehabilitation planning and contribute to meaningful evaluation of PC interventions.

Preliminary evaluation of the psychometric properties of the EFAT[4] indicated acceptable reliability and support for construct validity. The research also led to changes in some of the items. Thus reliability should be tested on the revised instrument.

Inter-rater reliability is the property of interest in this study. It is the extent to which administration of the same test by two or more people yields the same results. A reliable therapist is one who will be able to measure the same variables with consistent scores.[5] Without reliability, we cannot have confidence in the data we collect or draw rational conclusions from those data.[5]

In many studies, raters undergo a period of formal training prior to rating to ensure that their methods are standardized (formally trained). This is especially important when using a new instrument or when subjective observations are used. Even when standardized definitions are provided and the same performance is viewed,[5-8] different raters will not always agree on the scoring of the behavior in question.[5] Obtaining high inter-rater reliability by two trained test administrators is promising, but it is unrealistic to think that all future administrators will receive the same training or give the same attention to test administration protocols.

Like many instruments used in clinical practice, the EFAT was intended to be used by clinicians who teach themselves the administration procedures by reading the test manual (self-trained). A test manual containing standardized guidelines for administration has been developed to facilitate use of the EFAT.

The purpose of this study was to establish the inter-rater reliability of the revised version of the EFAT (EFAT-2) and to determine whether both formally trained and self-trained physical therapists (PTs) had an acceptable level of inter-rater reliability.

1.2 Limitations

Some limitations of the study include:

- Subjects - two independent groups of subjects were used. Comparison between the two pairs of raters (formally trained and self-trained) was not possible.
- Design – having both raters view the same performance was done to limit subject variability but allowed one PT to observe the other. The usual practice would be to see patients individually. PTs may become more reliable just by observing each other and becoming more alike in the test administration.
- PT behavior - PTs may have been more careful and thus may have higher reliability coefficients as a result of being in a study. This change in behavior may lead to an overestimation of the reliability of the PTs as used in the clinical situation.

1.3 Delimitations

This study is delimited to:

- PTs as raters: results may not be the same if other health professionals are used as raters.

- Evaluation of PC patients in continuing care: PC patients in continuing care facilities may not be representative of all PC patients e.g. those at home or those in acute care settings and results may be generalizable only to those PC patients involved.
- Exclusion criteria of patients with severe cognitive impairment or patients who do not speak English. The present study has provided no information regarding the use of EFAT-2 for these individuals.

CHAPTER 2

LITERATURE REVIEW

In recent years there has been increasing attention to issues related to quality of life (QOL) of people with cancer and a recognition of the potential for rehabilitation.[9] As the specialty of PC has grown so has the contribution of rehabilitation in that milieu.[10] As a result, more PTs are actively involved in rehabilitation of people with cancer during all phases of the disease.[9]

The issues to consider in the measurement of rehabilitation status in cancer care are numerous. They are as pertinent in the terminal stage of the disease as they are at other times in the cancer trajectory. Use of sensitive, valid and reliable measures in PC is crucial.[11] One reason for the few rehabilitation efficacy studies in PC may be the lack of appropriate measurement tools.

The EFAT is an outcome measure designed for use in PC. It was developed to describe and document the functional status of cancer patients throughout the palliative stage of their illness. The EFAT may be used as an assessment in everyday management of patients with terminal cancer. It evaluates the extent to which patients are able to interact and mobilize even when severely limited by their disease.

The following review of the literature covers existing performance measures in oncology and PC. Two outcome measures from the rehabilitation literature are also considered. The EFAT is then discussed in detail including its development and preliminary psychometric testing.

2.1 Performance Status Measures in Oncology and Palliative Care

A review of the literature revealed two performance status measures in oncology – the Karnofsky Performance Status scale[12] (KPS) and the Eastern Cooperative Oncology Group performance status scale[13] (ECOG). More recently the Palliative Performance Scale[14] has been developed. The KPS and the ECOG are widely used in oncology and PC. These scales are used extensively in research to classify patients into groups and to measure the outcomes of medical interventions. In spite of their common use there is limited information about the validity and reliability of these scales.

The KPS,[12] first developed in 1948, has since been used to measure functional status in cancer,[15] predict prognosis in terminal illness,[16] and determine outcomes in geriatric populations.[17] The KPS numerically describes in an easily administered, global score, the patients' ability to carry on normal activity and work, their need for a certain amount of custodial care, or their dependence on constant medical care. The KPS is an objective scale based on clinical assessment of health care providers. Inter-rater reliability coefficients between physicians and other health care providers have ranged from modest [r=0.34[3]] to good [r=0.69,[15] r=0.89[18]]. The highest inter-rater reliability occurred when raters were trained on the KPS.[17]

Similarly, the ECOG[13], provides a single score that describes the patients' activity from fully active (0) to completely disabled (4). Studies on the ECOG are sparse. Sorensen et al[19] established the reliability of the ECOG scale by measuring non chance agreement between three oncologists. The Kappa statistic for PS was 0.55, 0.48, 0.31, 0.43, and 0.33 for PS 0, 1, 2, 3, and 4 respectively. While overall chance agreement was moderate, agreement with regard to allocation of patients to PS 0-2

versus PS 3-4 was high. This is of note because this cut off is often used to select subjects according to performance status for clinical studies. Usually only subjects with PS 0-2 are included.

Clinical experience of the author[4] suggests that the scores on the KPS and ECOG for patients with terminal cancer tend to cluster at the low end of the scales. The "floor effect" limits their sensitivity to monitor change. Furthermore, the scales do not evaluate the functions assessed and treated in rehabilitation.

Anderson et al.[14] have identified a need to have an assessment tool which accurately reflects the changing physical status of PC patients. They have developed the Palliative Performance Scale, a new tool based on the KPS. The Palliative Performance Scale guides assessment of functional performance and provides a framework for measuring the progressive decline in palliative patients. The authors[14] feel that the KPS "suffers from being somewhat outdated." The emphasis on hospitalization does not accurately reflect the "closer to home" shift of resources in many health care systems today. The Palliative Performance Scale contains no reference to the location of care. Assessment variables are related to physical deterioration such as intake, mobility, and level of consciousness. By accurately reflecting physical change, the Palliative Performance Scale may have a role in prognostication, research, and program planning related to symptom control, drug costs, nursing, auxiliary requirements and respite needs.[14] The authors[14] stated that reliability and validity testing were underway, but to date no literature on the psychometric properties of the tool are available.

2.2 Additional Tools for Use in Oncology

The Functional Independence Measure (FIM)[20] from the rehabilitation literature has been documented as an outcome measure in oncology. Although the FIM has been used to evaluate functional outcome following rehabilitation of the cancer patient in the "pre-palliative" phase of the disease,[21,22] there is no documentation in the literature describing its use in a PC setting. Because of the poor physical condition of most patients with advanced cancer, many patients would require total assistance on some items. Therefore a consistently low score would emerge. This floor effect, reduces variability and ability to measure change.

In her review, Fulton[11] cites four standardized measures of rehabilitation status suitable for use in oncology: the Cancer Rehabilitation Evaluation System[23] and Cancer Rehabilitation Evaluation – Short Form,[24] the Edinburgh Rehabilitation Status Scale[25] (ERSS), the Rotterdam Symptom Checklist[26] and the Hospital Anxiety and Depression Scale.[27] Fulton indicates that a similarity between the measures is that they capture elements of QOL. She states that the measures may be used in studies of outcome or, additionally, as assessment tools in day to day management of patients with cancer. Their use may enable PTs to detect the needs of patients with cancer.

Of particular interest is the ERSS.[25] The ERSS was originally designed to measure changes which occurred during the course of disabling illness. The ERSS consists of 4 subscales, dependence/independence, activity/inactivity, social integration/isolation, and effect of symptoms on lifestyle. The scale is completed by a clinician or researcher who rates the patient on each subscale from zero to seven. The higher numbers in each case indicate a greater severity. Psychometric testing of the

ERSS was carried out by the developers of the scale.[25] It was shown to be sensitive to change over time and that it could be used as a simple assessment on the effectiveness of a rehabilitation program. The sample, although representative of many disabilities, did not include persons with cancer.

The original description of the ERSS[25] contained a comparison of the scale (n=30) with two other scales, the Barthel[29] and the PULSES.[30] A high degree of correlation for total scores was seen between all three scales. No examination was made though of individual components of each of the scales.

Using the Barthel as the gold standard, Mattison et al.[28] conducted a larger study with the ERSS. The purpose of the study was to examine the correlation between individual components of each of the above scales. The 3 scales were administered to 364 patients with primarily neurological diagnoses, and correlations between subscales were carried out. The results indicated that all three scales measured disability and all three were significantly related. Nonetheless, there were weak correlations between some of the individual subscales. The authors[28] suggested that further studies were required to evaluate the usefulness of the scales to give an overall picture of disability.

Fulton[11] used the ERSS to monitor the rehabilitation needs of women with metastatic breast cancer. She found a positive correlation between the ERSS total score and the global scores of the Cancer Rehabilitation Evaluation System – Short Form.[24] Use of the ERSS in a palliative setting has not been documented.

2.3 Development of the Edmonton Functional Assessment Tool

No single comprehensive instrument was found to describe the functional status of patients with terminal cancer. Thus, development of an instrument to meet the needs of PC patients was begun.

Recommendations by Cella and Tulsky[31] for measurement of patients with cancer were considered in developing the EFAT. Their recommendations included a focus on the patient's abilities, provisions for repeated assessments, and communication of concise and relevant information to the healthcare team. They also suggested that a measure should contribute to patient care without posing a burden to patients, families or staff, and should still be sensitive enough to meet the needs of research. As well, a rehabilitation perspective of functional status within the domains of impairment and disability as defined by the World Health Organization[32,33] served to guide the developers.

Functions assessed in PC were chosen from the activities observed to be important to patients, even in the terminal stage. Ten key functions were selected by the developers in consultation with five therapists working at three acute care facilities, one home care facility and one long term care facility in Edmonton, as a starting point for the development of an instrument to evaluate functional status in PC.

The EFAT consists of two parts. Part 1 contains ten items describing symptoms/functions the patients may experience or are required to perform. These include communication, mental alertness, pain, sensation, respiratory function, balance, mobility, activity level, wheelchair mobility and ADL. Part 2 is a single item that summarizes the PS after the 10 specific functions are assessed.

Four descriptors are used to score each item. Each descriptor outlines the behaviors that the patient must perform to be awarded a rating from 0 to 3. A" 0" rating represents independent functional status and a rating of "3" represents a total loss of functional status.

2.4 Preliminary testing of the EFAT

The first version of the EFAT has been developed and tested.[4] Tests for inter-rater reliability and concurrent validity were conducted using a sample of twenty-five inpatients on the palliative care unit (PCU) at the Edmonton General Hospital (EGH). They were evaluated on two separate occasions by two raters. Inter-rater reliability of the EFAT expressed as an intra-class correlation was established at 0.88. Concurrent validity of the EFAT was demonstrated by correlating the total EFAT score with other performance measures in the field, namely the KPS and ECOG. The Pearson Product Moment Correlations were $r = -0.79$ ($p = 0.0001$) and $r = 0.85$ ($p=0.0001$) respectively. The total EFAT score was also strongly related to its accompanying global PS rating scale at $r = 0.90$ ($p =0.0001$).

Gathering evidence to support the validity of the construct of functional status began with the preliminary testing of two hypotheses. The first hypothesis was that the EFAT scores would increase from admission to discharge, and portray the deterioration of function expected with the progression of the disease. The mean discharge EFAT scores were significantly higher than the mean admission EFAT scores for the total sample of patients [$t = -3.216$, $p = 0.01$].

The second hypothesis stated that the mean EFAT scores would be higher at both admission and discharge for patients who remained on the unit or were transferred

to another unit (n = 88) than for the patients who were discharged home (n = 13). Patients in the home group showed significantly lower mean admission scores on the EFAT ($\bar{x}$ = 26.46) than patients in the unit group ($\bar{x}$ = 40.78) [t for unequal samples = 3.25, p=0.0046]. The difference in EFAT scores from initial to final assessment was significantly greater for the unit group (n = 53), [t = -7.509, p = .0001], but not for the home group (n=7), [t = 0.072, p= 0.94]. The findings provided initial evidence that the EFAT distinguished between the functional status of these two groups.

Although analysis of preliminary data collected on the EFAT has lent support to inter-rater reliability and has added information to support the concurrent and construct validity of the tool, limitations of the study have been identified. First the raters involved in the validation study were involved in the development of the tool. Second, the sample size (n=25) for concurrent validity was small. Third, the groups used to provide evidence for construct validity were unequal. The results of this preliminary study suggested that the EFAT required further research and development, but showed potential to evolve as a useful clinical tool in palliative care.

2.5 Testing of EFAT-2

A number of recommendations for the development and testing of the EFAT-2 were made by the authors.[4] The descriptors for pain, mental alertness and respiratory function were reworded to clarify behaviors to be observed. The balance item was redefined to include either sitting or standing balance. The wheelchair mobility and activity items were collapsed and a new item called locomotion was devised. The sensation item was deleted because a sensory impairment did not translate into a functional problem. Two new items thought to impact functional status - fatigue and

motivation – were included. The items included in the EFAT-2 are outlined in Appendix A. Guidelines for scoring were clarified (Test Manual - Appendix B).

The EFAT-2 has been used as part of the rehabilitation database on the acute PCU at the Grey Nun's Community Hospital & Health Center (GNCHC) since January 1994. Data was collected on 277 admissions to the acute PCU from January 1994 to June 1996.[34] The EFAT-2 was administered on admission and then approximately weekly until death, transfer to a continuing care PCU or another facility, or discharge home. The data was statistically analysed to provide further information on the psychometric properties of the tool.

Principal components factor analysis was performed for the EFAT-2 items. Eigenvalues of greater than 1 were used to select an appropriate number of components to be derived. Using a 3 factor oblique solution, communication, dyspnea, fatigue, mental status, and motivation loaded on factor 1 (.58 to .86); ADL, balance, locomotion, and mobility loaded on factor 2 (.60 to .97); and pain loaded on factor 3 (.99). As has been noted previously,[4] pain was an independent item. Factor 2 includes various aspects of physical ability while Factor 1 includes non-physical components. The degree of internal consistency of the ten items was determined by calculating Cronbach's alpha, which was 0.871.

A one way analysis of variance (ANOVA) was used to determine whether there was a difference in EFAT - 2 total score according to discharge location (Group 1-died on unit, Group 2-transfer to continuing care PCU, Group 3- discharge home, Group 4-other). The result was significant [$F(3, 273)\ 20.5963$, $p < 0.000$ two tailed test]. A post-hoc Newman-Keuls test indicated that group 1 was significantly different from the

other three groups. There was also a significant difference between Groups 2 and 3. The mean total score for each group was: Group 1 = 16.1, Group 2 = 13.5, Group 3 = 9.0, and Group 4 = 10.5. These findings provided further support for the construct of functional status and suggested that the EFAT-2 distinguished between the functional status of these four groups.[34]

The EFAT was designed in an attempt to alleviate some of the shortcomings of existing PS measures used in PC. Preliminary research has led to the inclusion of new items and revision of some of the existing items (EFAT-2). Methodological standards in the development of assessment tools require that psychometric properties be re-established.

2.6 Reliability

The concept of reliability at first glance is straightforward. According to Streiner and Norman,[35] before one can obtain evidence that an instrument is measuring what it is intended, it is necessary to gather evidence that the scale is measuring something in reproducible fashion. The first step in providing evidence of the value of an instrument is to demonstrate that measurements of individuals on different occasions or by different observers, or by similar or parallel forms produce the same or similar results.[35]

Dumholdt[36] defines reliability as the "degree to which test scores are free from errors of measurement." Her introduction to theories, components and measures of reliability is discussed below.

2.6.1 Measurement Theories

There are two basic concepts in measurement theory; classical measurement theory and generalizability theory. In classical measurement theory, the assumption is

that every measurement or obtained score contains a true component and an error component. Additionally, each individual has a single true score on the measurement of interest. Because we can never know the true score for any measure, the relationship between repeated measures is used to estimate measurement errors. A measurement is said to be reliable if the error component is small, therefore allowing consistent estimation of the true quantity of interest. With classical measurement theory, all variability within a person's score is regarded as measurement error.[36]

Generalizability theory recognizes that there are different sources of variability for any measure. Measurements are studied in ways that permit the researcher to divide the measurement error into sources of variability or facets. According to Dumholdt,[35] the generalizability approach appears to have applicability to the study of measurement in PT because it acknowledges and provides a way to quantify the many sources of variability that PTs see in their patients from day to day.

2.6.2 Components of Reliability

There are a number of ways in which reliability measures may be obtained, and the size of the reliability coefficient will be a direct reflection of the particular approach used.[35] Various components of reliability are outlined below.

Instrument Reliability: There are 3 broad categories of PT measurement– biophysiological, self-report, and observational.[36] Different instruments are used for different measurements. The method used to determine an instrument's reliability depends on the type of instrument. Examples of instrument reliability are test-retest (the assessment of scores on two or more administrations of the test), parallel forms (similar forms of the test are each administered once), split half (portions of the test are

compared with each other), and internal consistency (responses to individual items are evaluated).

Intra-rater reliability is "the consistency with which one rater assigns scores to a single set of responses on two occasions."[36(p.155)] Inconsistency may be the result of rater error or inconsistent subject performance.

Inter-rater reliability is "consistency of performance among different raters or judges in assigning scores to the same objects or responses. It is determined when two or more raters judge the performance of an individual or a group of subjects at the same point in time"[36(p.155)] This is a pure measure of inter-rater reliability. When subjects are viewed at two different times, variability may be attributed to rater differences or subject variability.

Intra-subject variability refers to the actual changes in subject performance from time to time. Some measurements in physical therapy may appear to be unreliable because the occurrence being measured is intrinsically variable eg. spasticity.[36] Obtaining a pure measure of intra-subject variability is impossible unless one has a perfectly reliable instrument and a perfectly reliable rater. Therefore, most test-retest reliability coefficients contain a degree of instrument error, rater error and true subject variability.

2.6.3 Quantification of Reliability

Reliability is quantified in two ways, relative or absolute.[36] Relative reliability looks at the relationship between two or more sets of repeated measures. It is measured with some form of a correlation coefficient, which indicates the degree of association between repeated measures of the variable of interest. A reliability coefficient is usually

a ratio of variability between individuals and the total variability in scores. The reliability is expressed as a number between 0 and 1 with 0 representing no reliability and 1 representing perfect reliability.[35]

Absolute reliability is the extent to which a score varies on repeated measurement.[36] The statistic used to measure absolute reliability is the standard error of measurement (SEM). It is important to know how much variability in the scores could be expected because of measurement errors. Because reliability coefficients and SEMs provide different views of reliability, it is important to document them both.

There is ongoing debate over what constitutes an acceptable level of reliability. Several authors[8,37,38] suggest as a general guideline that values above 0.75 are indicative of good reliability and those below 0.75 are poor to moderate. Streiner and Norman[35] argue that reliability cannot be conceived of as a property that a particular instrument does or does not possess; rather, any measure will have a certain degree of reliability when applied to certain populations under certain conditions.

2.7 Summary Statement

As more people live longer with cancer due to improved medical treatment, physical therapists will be involved in their rehabilitation. As the PC specialty grows it becomes imperative to use standardized measurement tools to evaluate and document the effectiveness of PC interventions. Little published information is available regarding measurement in this area. The EFAT was developed in an attempt to alleviate some of the shortcomings of existing performance status measures used in PC. Preliminary psychometric testing of the EFAT led to a revision of some of the

individual items of the tool. Establishing the reliability of the EFAT-2 is the next step in its development.

CHAPTER 3

METHODOLOGY

This study was designed to establish the inter-rater reliability of the EFAT-2 and provided an opportunity to determine whether both formally trained and self-trained PTs obtain an acceptable level of inter-rater reliability.

3.1 Objectives

The objectives were:

1. To determine the inter-rater reliability of formally trained raters using the EFAT-2
2. To determine the inter-rater reliability of self-trained raters using the EFAT-2

3.2 Hypotheses

The hypotheses were:

1. There will be a high (>0.80) intraclass correlation (ICC) for formally trained raters using the EFAT-2.
2. There will be a high (>0.80) ICC for self-trained raters using the EFAT-2.

3.3 Raters and Subjects

The raters for the study were PTs who met the following criteria:

1. A minimum of one year clinical experience, six months in palliative care
2. Minimal exposure to the EFAT preferred

Four physical therapists meeting the above criteria volunteered to participate – a non random sample of convenience. The PTs functioned in pairs; one pair (R1, R2) worked at the Edmonton General Hospital (EGH), and the other two PTs (R3, R4) worked for either the Edmonton Homecare Program or the acute PC Program at GNCHC.

The subjects constituted a non-random sample of convenience. There were two independent samples; sample 1 contained subjects exclusively from the PCU at the EGH, sample 2 contained subjects from the PCUs at EGH, St. Joseph's Auxiliary Hospital and GNCHC. R1and R2 assessed sample 1 and R3 and R4 assessed sample 2.

Patients from the PCUs were included if they:

- had a life expectancy of at least one month
- were comfortable and willing to participate, and
- provided informed signed consent prior to entering the study

All subjects admitted to the study were:

- experiencing progressive disease where the focus of care was on comfort, and improving quality of life, not cure,
- requiring active care to alleviate distressing symptoms related to physical, psychosocial and spiritual needs.

In addition, patients on continuing care PCUs:

- could not be managed at home
- did not require acute/tertiary care
- had an expected length of stay of approximately 2 months
- were over 18, and
- had accepted a no code status

The subject criteria were similar to the admission criteria for continuing care PCUs as set by the Edmonton Regional PC Program (Appendix C). According to data collected by the Edmonton Regional Program,[39] patients on the three Edmonton continuing care PCUs did not differ according to age, gender, or diagnosis. Functional

status was not consistently monitored. The demographics of the two groups were compared statistically at the completion of the study.

Patients signed informed consent sheets to participate in the study. Patients with questionable competency or with limited understanding of the English language were excluded from this study. Clinical judgment of the PI was used to establish competency for informed consent. Clinical judgment of the raters was used to determine whether the subjects would be able to perform the EFAT-2. The mini mental state exam (MMSE-Appendix D) was administered just before or soon after the EFAT-2 to document the mental status of the subjects at the time of EFAT-2 administration.

3.4 Sample Size

Based on an alpha level of 0.05, and power of 80%, sample size was 36 for each group of subjects (Appendix E). The sample size was based on a minimum acceptable ICC value of 0.80. Accrual continued on a consecutive basis for each sample from day 1 of the study until the sample size was achieved.

3.5 Enrollment

There was a maximum of 26 patients on the continuing care PCU at the EGH. The unit averaged 10 admissions per month. Based on information from previous work (personal communication Dr. E Bruera) it was estimated that consent would be obtained in 75% of cases. Consequently it was anticipated that 7 patients would be enrolled each month and that data collection would be complete in 4.5 months. Accrual for sample 1 commenced 29 September 1997 and was complete by 5 March 1998. No subjects were accrued between 8 December 1997 and 9 January 1998.

Enrollment for the second sample of subjects occurred at three sites: EGH, St. Joseph's Auxiliary Hospital and GNCHC. This was done in an attempt to facilitate accrual as data collection for the second sample was delayed until 3 January 1998. There were 21 subjects from St. Joseph's Auxiliary Hospital, 8 subjects from GNCHC, and 7 subjects from EGH. Data collection was complete by the 7 May, 1998.

The PI was responsible for communicating with the unit managers regarding potential subjects for the study. The PI explained the purpose and methodology of the study to interested patients. Those who agreed to participate were asked to read and sign an informed consent form (Appendix F). As patients were admitted to the study they were assigned a study number to ensure confidentiality.

3.6 Study design

The study addressed the issue of reliability. A pair of formally trained PTs and a pair of self-trained PTs were used. Two similarly trained raters used the EFAT-2 to simultaneously assess consenting subjects on one occasion. The raters' scores were statistically analysed to provide a reliability coefficient.

The PI conducted formal training. It consisted of:

- raters reading the test manual
- raters and PI discussing the guidelines for administration
- raters and PI scoring 3 simulated case scenarios and
- clinical application of the EFAT-2 by the raters and the PI simultaneously

Once the raters were rating total score within 2 of the PI they were deemed accurate and data collection was begun.

The self-training consisted of:

- raters reading the test manual and teaching themselves to administer the EFAT-2
- raters administering the test to two or three patients

Once this was accomplished, data collection began.

All raters were asked to administer the EFAT-2 as per instructions for use in the test manual (Appendix B). They were also provided with an addendum to the guidelines clarifying scoring criteria of pain, mental status, dyspnea, fatigue, motivation, and PS (Appendix G).

Patients were admitted into the study once their symptoms were controlled and they provided written consent. In general, subjects were assessed during the same week that they were admitted to the study.

One rater administered the EFAT-2 while the other rater observed. Both raters scored the EFAT simultaneously and independently of each other. Because the administrator of the EFAT can influence the test, raters were randomized as administrator or observer for each patient prior to the initiation of the study. This controlled for potential rater bias, that is, each rater had the opportunity to generate a subject-rater relationship.[8] The administration of the total EFAT-2 took approximately 15 - 20 minutes.

In the preliminary psychometric studies on the EFAT,[4] inter-rater reliability was assessed by having two raters evaluate patients within a 24 hour time period of each other. Although reliability was acceptable, percentage agreement of the raters on some items was moderate only. One possible explanation was that this patient population

may show significant change in a 24 hour period. To control for the effects of time, subjects were seen simultaneously in this study.

Terminally ill cancer patients are patients in whom multiple cancer treatments such as surgery, radiotherapy, and chemotherapy, have failed and in whom an aggressive cancer treatment is no longer an option. These patients are affected by a number of symptoms: pain, nausea, and fatigue to name a few. From clinical experience the author had concluded that these patients are less fatigued in the morning and more likely to feel like moving when their symptoms are controlled. Comfort of the subjects was ensured by arranging the assessments in the morning and in conjunction with their regular medication whenever possible.

3.7 Data Collection

The PI was responsible for recruiting subjects for the study. A record sheet was completed for each patient who agreed to be involved. The record sheet contained space for identifying the hospital number of each patient, the study number given each patient on admission to the study and demographic data.

Data collection from the EFAT-2 was the responsibility of the raters. Scores on each item of the EFAT-2 were recorded on an EFAT-2 record sheet. Each rater placed completed EFAT-2 records in a designated mail box (R1, R2) or envelope (R3, R4). Completed record sheets were collected by the PI and stored in a locked file cabinet at the University of Alberta. Raters were asked not to discuss the scores until the data collection for the entire study was complete. Subjects were asked to refrain from discussing their performance with the raters. The PI provided feedback to the subjects after the EFAT was administered.

3.8 Statistical Analysis

All data were analyzed using an alpha level of 0.05. Appropriate descriptive statistics were used to describe the two independent samples on basic demographics. The samples were compared using the following analyses.

Variable	Data	Descriptive Statistics	Inferential Statistic
Age	Interval	Mean and standard deviation	t test
MMSE	Interval	Mean and standard deviation	t test
Gender	Nominal	Frequency	X^2
Diagnosis	Nominal	Frequency	X^2

The following statistics were used to evaluate inter-rater reliability of the EFAT-2 for both formally trained and self-trained raters.

Variable	Data	Descriptive Statistic	Inferential Statistic
EFAT-2 total score	Interval	Mean and standard deviation	ICC (1,1) SEM
Individual item score	Categorical		Kappa

The ICC was calculated:

$$ICC = \frac{MS\ between - MS\ within}{MS\ between + (k-1)MS\ within} \qquad SEM = \sqrt{MS\ within}$$

The minimal acceptable level of the reliability coefficient in this study was 0.80. Confidence intervals were calculated for the ICC and allow us to comment on the range of the ICC with a desired level of confidence.[40] (Appendix H). The standard error of measurement (SEM) between the two raters was calculated. It reflects the extent of

expected error in different raters' scores.[41] It allows each subject's score to be considered with an error term.

A secondary analysis on the individual test items was performed. The Kappa statistic, a chance corrected measure of agreement has been used. It is possible to have good agreement on the total score but not on individual items.

3.9 Ethical considerations

Consent from the Ethics Committees of the Caritas Health Group and St. Joseph's Auxiliary Hospital was obtained. The PI approached potential participants and explained the purpose and methodology. Information about the study was outlined in the consent form (Appendix F). Consenting patients were asked to sign the consent form that outlined right to withdraw, confidentiality, and the risks and benefits involved.

In the study the subjects' normal functional level was assessed. Subjects were not expected to do any activity they were not capable of or would feel unsafe trying. One potential risk was fatigue and subjectsts were allowed to rest whenever they wanted to.

PC patients are a vulnerable population. They often suffer from devastating physical and psychosocial symptoms. Bruera et al [42] found that approximately 80-90 % of terminally ill patients who were approached agreed to participate in research. They found that patients and families would participate if they understood the purpose of and potential benefit from the study. Understanding and cooperation of subjects was fostered in the present study by:

- use of simple language and a concise consent form. This took into account the physical and mental fatigue of patients as well as the psychological distress which may have been present
- assessing mental status prior to obtaining informed consent,[42,43]
- providing adequate time for patients and families to read and discuss the consent form, and
- ensuring the raters were flexible; able to recognize if the patient was becoming fatigued or confused and were able to explore options of stopping for a rest or stopping completely.[43]

Participation in the study provided no personal benefit to the subjects. However, it was hoped that in the long term, adequate assessment and documentation of functional status of PC patients would lead to effective treatment of PC patients. Some patients have found it worthwhile to participate in research in the hope of contributing to the well being of others who may experience a similar illness.[42-44]

All signed consent forms and collected data were kept confidential. A copy of the ethics approval was kept on file and a copy was provided for the manager of all units involved. Each subject was informed as noted above and received a copy of the signed consent. Original consent forms became part of the subject's hospital chart.

3.9 Focus Group

A focus group[45] was held six weeks after the last subject was rated. The PI, three of four raters (R1, R2, and R3), and the PI's supervisor were in attendance. The purpose of the focus group was to determine whether the PTs who had volunteered as raters considered the EFAT-2 a clinically useful tool.

An explanation of the interview process was provided and all group members were asked for their contributions. Raters were assured that their answers would not be considered "right" or "wrong" and that the data from the discussion would not be published in a form that would threaten the anonymity of them. Written and Audio records of the discussion were kept. The raters were encouraged to speak with each other rather than to the PI.

The raters were given three basic tasks:

- To describe their overall impression of the EFAT
- To describe any items they had difficulty rating, and what made them difficult
- To describe the item(s) in the tool they found redundant or that did not provide clinically useful information

At the end of an hour no new areas for discussion were brought up and the PIs impressions of the discussion were clarified with the group. The PI later transcribed the tape.

3.11 Summary Statement

This study was designed to determine the inter-rater reliability of a revised version of the EFAT (EFAT-2). It also provided the opportunity to determine whether both formally trained and self-trained PTs obtain an acceptable level of reliability. Ethical approval was obtained from the Caritas Health Group and St. Joseph's Auxiliary Hospital. Two samples (n=36 each) were accrued from three healthcare facilities in Edmonton.

The EFAT-2 was administered to consenting PC patients by one of two independent PT rater pairs. One pair was self-trained and the other pair formally trained.

An intraclass correlation was used to establish a reliability coefficient for each similarly trained pair. The SEM and a 95% CI were reported. The Kappa statistic was used to determine agreement between similarly trained raters on individual items.

CHAPTER 4
RESULTS

The results are presented in the following way: the characteristics of the physical therapist raters are given followed by the demographics of the two samples. The section is completed by presenting the reliability of the rater pairs for the total EFAT score and for individual items of the EFAT.

4.1 Characteristics of the Raters

The characteristics of the PTs who volunteered as raters are given in Table 1.

TABLE 1 Characteristics of the Physical Therapist Raters

Variable	R1	R2	R3	R4
Gender	Female	Female	Female	Female
Year graduated	1978	1970	1972	1991
Experience (yrs)	20	>25	>20	7
Post Graduate Education	M.A. (Soc.)		MScPT. (cand.)	
Training in EFAT-2	Self	Self	Formal	Formal

4.2 Characteristics of the Subjects

Table 2 presents descriptive data for the subjects. Although the two samples differed in age, they were similar regarding gender, diagnosis, and cognitive status. Information on patients who did not participate in the study is included in Appendix I.

TABLE 2 Characteristics of Subjects

Variables	Sample 1 (R1,R2)	Sample 2 (R3, R4)	P Value*
Age (yrs)	73.4±9.5	67.5±13.7	0.040
MMSE	24.6±5.7	26.1±3.72 (n=33)	0.195
Age Related Norm	24.2±3.2	24.5±5.3	
Gender			0.475
Male	17	14	
Female	19	22	
Diagnosis (Cancer)			0.655
Gastro Intestinal	13	12	
Genito Urinary	5	9	
Lung	10	6	
Breast	6	6	
Other	2	3	

* of t test for interval data and X^2 for categorical data

4.3 Reliability

Table 3 contains the intraclass correlations (ICC) of the rater pairs for total score on the EFAT-2. The SEM and 95% confidence intervals (CI) are also included. The values were calculated from results of repeated measures one way ANOVAs shown in Table 4. The results suggest that there is excellent inter-rater reliability between the two similarly trained raters.

TABLE 3 EFAT-2 Total Score (mean ±SD) for Four Raters with ICC and SEM for each Pair of Raters

	Mean ± SD	ICC	CI	SEM
Rater 1	12.19±6.61			
		0.97	(0.97, 0.99)	1.09
Rater 2	12.28±6.44			
Rater 3	16.11±6.86			
		0.95	(0.95, 0.99)	1.44
Rater 4	14.81±6.36			

TABLE 4 Analysis of Variance of EFAT-2 Total Scores for Sample 1 and Sample 2

	Source of Variation	SS	df	MS	F	P
Sample 1	Between People	2937.49	35	83.93		
	Within People	43.50	36	1.21		
	Between Measures	.13	1	.13	0.1009	0.75
	Residual	43.38	35	1.24		
Sample 2	Between People	3006.38	35	85.89		
	Within People	75.50	36	2.09		
	Between Measures	30.68	1	30.68	23.9588	0.00
	Residual	44.82	35	1.28		

The Kappa values and percentages of agreement within the rater pairs on individual items are presented in Table 5. Percentage of agreement is the number of subjects on which the raters agreed as a proportion of all subjects studied. The average percent of agreement on individual items for the formally trained raters was 82% and for the self- trained raters 85%. Items with agreement lower than 80% were motivation for both pairs of raters, and fatigue, mental status, and PS for the self-trained raters.

The average percent disagreement for the self-trained and formally trained raters was 15% and 18 % respectively. Items where disagreement spanned more than one rating point for the self-trained raters were balance, motivation and PS and for the formally trained raters were communication, locomotion, mental alertness, mobility, motivation, pain, and PS. Simple overall agreement on the EFAT items is high. However, we would expect a certain amount of agreement to occur by chance alone.

TABLE 5 Kappa values and Percentage of Agreement for Rater Pairs on Individual Items of the EFAT-2

EFAT ITEMS	K R1,R2	% AGREE	K R3,R4	% AGREE
Activities of Daily Living	.704	80.5	.671	86.1
Balance	.844	88.8	.945*	97.2
Communication	.250	80.5	.512	88.8
Dyspnea	.721	80.5	.955	97.2
Fatigue	.916	94.4	.656	77.7
Locomotion	.883	91.6	.805*	88.8
Mental Status	.708	83.3	.557	77.7
Mobility	.962	97.2	.886*	94.4
Motivation	.528	66.6	.172	36.1
Pain	.826	88.8	.849	88.8
Performance Status	.736	80.5	.574*	75.0

* calculated by hand

The advantage of the Kappa coefficient is its correction or "adjustment" for the amount of agreement that can be expected to occur by chance alone. Kappa values are available for each item in the EFAT-2 for R1, R2. For R3, R4, Kappa was calculated for 7/10 items. In the remaining 3 items; balance, locomotion, and mobility and for the PS item, the statistical program would not compute a Kappa value because not all categories were used by both raters. In the case of balance, locomotion and mobility, R4 did not utilize the "2" rating and in PS, R3 did not use the "0" rating. The Kappa statistics for the balance, locomotion, mobility and PS items were calculated by hand.

4.4 Focus Group

The focus interview provided insight into the utility of the EFAT-2 and identified items for review. Overall the raters found the EFAT-2 easy to administer and quick to use once they were familiar with the guidelines. The EFAT-2 did not tire the subjects and was thought to be easy enough to use for reassessment. It provided information from a rehabilitation perspective in a concise format on a single page. Although the EFAT-2 was not designed to prescribe treatment, the raters felt it provided enough information on which to base a treatment or intervention.

The raters described 3 items as challenging to rate: motivation, dyspnea, and fatigue. In some items the difficulty was experienced by the raters themselves and in others the raters perceived that the subjects had difficulty understanding what was being asked of them. The items will be addressed in turn.

Motivation was identified as difficult to rate because the raters were unclear on what the item was actually measuring. The confusion lay between recognizing the emotion or desire, which incites action from within an individual, and the physical

capacity to carry out the action. The formally trained raters in particular found it difficult to rate this concept when meeting the subjects for the first time.

Explaining the concept of the dypsnea rating scale was difficult for two of the raters and from their perspective proved hard for some subjects to understand. Therefore they were not always certain the rating they gave when using the dyspnea rating scale was valid. When the scoring was based on the use of oxygen, it was far more objective.

All raters related difficulty scoring the fatigue item. Subjects were asked " Do you feel tired?" When their answers correlated with what the raters saw during the assessment, the item was easier to rate. If the subjects answer did not correlate with observed behavior during assessment, the raters experienced difficulty scoring the item. The self-trained raters used the nursing notes to validate their ratings. The meaning of "rest" and "daytime" were unclear to the raters. They perceived some difficulty with using the percentage descriptors because of this.

Two items were identified as redundant or not providing clinically useful information. First, all the raters felt the motivation item did not provide clinically useful information, although they all felt that motivation was an integral part of rehabilitation. Second, one of the rater pairs did not use the "0" or "1" rating for ADL. Hence they felt the ADL descriptors lacked the sensitivity to provide clinically useful information.

Scoring of the performance status item was done differently by each pair of raters. R1 and R2 looked at the subjects globally and placed each subject in the category where they met the majority of the attributes as per the current guidelines for administration. During formal training it was established that R3 and R4 would place the subject in the highest category which contained an attribute that described the subject. For example if a

subject met all the criteria for category 2 in PS except that they were dependent for ADL, the formally trained raters would have rated them as a "3". In contrast, the self-trained raters would have rated them as a "2".

The raters noted that subjects would, on occasion, be bed bound or would choose not to move about during the EFAT-2 administration even though they might do so on a regular basis. In either case the raters would score the individuals as "3" on balance, mobility and locomotion according to the instructions in the test manual. The raters identified the scoring as misleading because one cannot tell whether the individual received a rating of 3 because they were unable to perform the task or because they chose not to do the task. From the raters' perspective, if a subject chose not to do the task, the "3" rating did not give an accurate portrayal of the individuals usual functional status.

CHAPTER FIVE
DISCUSSION

5.1 Reliability

The results of the present study support the original hypotheses which stated that there would be a high (>0.80) ICC for both the formally trained raters and the self-trained raters using the EFAT-2. The narrow range of the CI for the ICCs and the low SEMs support good inter-rater reliability. The Kappa values on 9/10 items of the EFAT-2 suggest acceptable to high levels of rater agreement on the individual items. It appears that both formally trained and self-trained raters may obtain an excellent level of reliability with the EFAT-2.

The CI allows one to report with confidence on the strength of the reliability coefficient and provides a unitless estimate of the reliability of measurement. An illustration of the derivation of the CIs is shown in Appendix H. In the present study, the CIs allow us to be certain that the true magnitude of the ICC falls between 0.97 and 0.99 for R1, R2, and between 0.95 and 0.99 for R3, R4 95% of the time. The CIs define a very narrow range for the ICCs supporting the high reliability.

The precision of measurement is estimated by the SEM.[7,41] It is expressed in the same units as the original measurement. The interpretation of the SEM depends on the type of reliability coefficient used to calculate it. When the reliability coefficient reflects inter-rater reliability, as in the present study, the SEM indicates the amount of expected error in each raters score. Each single measurement obtained is associated with a degree of error that has a known probability.

The SEM is really a standard deviation and can be used to form a band or CI around a given score. Because the SEM is based on a normal curve, we can be certain

that the CI formed by the observed score ±1 SEM would contain the true score 68% of the time. For example, if R1 (SEM= 1.09) administered the EFAT-2 and obtained a total score of 14, the true score would be contained within the interval 12.91-15.09 68% of the time. Similarly we can be certain that the true score would be contained within the range 11.82-16.18 (±2 SEMs) 96% of the time. The SEM is useful for a PT because in order to make a meaningful statement about whether a subject's condition has changed, the PT must know how much variability in the total score could be expected due to measurement error and how much variability is a clinically significant change.

It is important to provide both an estimate of reliability (relative reliability) and precision of measurement (absolute reliability).[41] A low estimate of relative reliability may not be of concern if a small SEM suggests that any inconsistency of measurement occurs in an acceptably small range. On the other hand, a relatively high ICC may not reflect an acceptable measurement if the SEM suggests that the precision of measurement is not acceptable for the intended purpose.

The original research on the EFAT[4] provides an illustration of a high ICC that may not be indicative of an acceptable level of measurement. The ICC for inter-rater reliability was 0.88 and was interpreted as being excellent. However, the original sample was more heterogenous than the present samples, and the variety in the sample may have artificially inflated the ICC. The heterogeneity was reflected in the large SD for both raters on total score. A large systematic bias was demonstrated by a significant difference in means of the raters on total score. The original data was further examined. A CI for the reliability coefficient and the SEM were calculated. The 95% CI for the ICC was 0.86 - 0.99, broader than those in the present study. The SEM was 7.10

suggesting the precision of measurement may have been questionable. It appears that the ICC, CI, and SEM in the present study support an improvement in reliability over the original EFAT.

Based on the Kappa values, the items of the EFAT-2 could be reliably assessed with the exception of motivation in sample 2. According to Landis and Koch,[37] Kappa values at or above 0.50 are considered as having a good level of agreement. In the present study, the majority of Kappa values range from 0.53 - 0.96. The Kappa values for communication and motivation are exceptions and are discussed separately later. In the original tool,[4] five items had Kappa values of less than or equal to 0.50 with the remaining items ranging from 0.61 - 0.81. In the present study not only have the Kappa values improved, a greater proportion of them are above 0.50.

Three main reasons may account for the improvement in reliability from the original EFAT: the method of test administration, the changes in the EFAT items and the clarification of the guidelines for administration. The rational behind each issue will be addressed separately.

Tests for inter-rater reliability of the original version of the EFAT were conducted using a sample of twenty-five subjects. Two independent raters assessed the subjects within 24 hours of each other. When observations are made at two separate times the variability attributed to the raters and the variability within subjects cannot be separated. In the present study, the raters observed the same performance. This has the potential to increase reliability.[5,36] When two raters simultaneously observe and rate a performance, the comparison between their scores is a purer measure of inter-rater reliability because they observe the exact same behavior. This eliminates a change in

the subject as a source of measurement error when comparing raters' scores.[5] Hence, reliability may increase.

When more than one rater is used to administer a test, the potential for differences in administration exists and can impact reliability. In the present study the raters were randomized to be either administrator or observer for each subject prior to commencing the present study. The PTs may have become more reliable by observing each other and by becoming more alike in the test administration.

Changes made to some EFAT items followed recommendations arising from preliminary research[4] and may have enhanced the reliability of the EFAT-2. The descriptors for pain, mental alertness and respiratory function were reworded to clarify the behaviors to be observed. The pain item was retained even though it performed poorly in terms of percentage of agreement, inter-rater reliability and item correlation with the whole EFAT. It was reworded in behavioral terms to ascertain its impact on performance of functions important to patients. This is in line with patient derived health domains where pain is usually described in terms of its impact on function ie: "it stops me from doing things".[45] The Kappa values of both rater pairs in the present study reflect excellent reliability on this item. The balance item was redefined to include either sitting or standing balance. The wheelchair mobility and activity items were collapsed and a new item called locomotion was devised. The sensation item was deleted because a sensory impairment did not translate into a functional problem. Two new items which were thought to impact functional status of PC patients, fatigue and motivation, were included. The overall aim was to enhance objectivity and thus, reliability of the tool may have increased.

Guidelines for use of the EFAT-2 were established subsequent to the preliminary research on the EFAT. They served to provide more consistent administration of the test. Instructions were also made clearer by asking that all items be evaluated. The total score on the EFAT-2 is 30. No provision was made to consider an item "not applicable" or "not evaluated." The 3 rating was to be used if an item could not be assessed or if it was unsafe to do so. An addendum to the guidelines, established prior to commencing the present study, established criteria to provide consistency in rating subjective items: pain, dyspnea, fatigue, and motivation. Rehabilitation staff using the EFAT-2 on the acute PC unit had identified these items as requiring further explanation. Suggestions for use in rating cognitive status were also provided. Both the guidelines and the addendum may have contributed to the increase in reliability found with the EFAT-2.

Many authors recommend observer training to reduce error variance. The specific strategies to use in training are usually unspecified.[5,35] According to Dumholdt,[36] the level of standardization in the measurement protocol may also impact reliability. The degree of standardization is the number of sources of variability within a reliability component that are controlled. Standardization may range from none at all to use of a highly structured protocol. In the present study, the formally trained raters followed a partially standardized approach. They were educated in the administration methods to be used with the EFAT-2. Sources of variation that were not controlled for included the time of day the EFAT-2 was administered and the time of the last medication. The self-trained raters followed a non-standardized approach. It was

anticipated that the formally trained raters might have had a higher reliability coefficient than the self-trained raters.

Because two separate samples were used it is not legitimate nor the intent of this study to compare the reliability of the pairs of raters. However it is interesting to note the circumstances of each pair, and how these might affect reliability.

The excellent reliability of the self-trained raters could have been due to a number of similarities between the raters:

- Although neither had used the EFAT-2 clinically prior to the study, both were familiar with the tool in that it had been developed in the same facility that they worked in.
- They were employed in the facility where they rated subjects, and one of them worked on the PC unit. They were perhaps more familiar with patients on the unit.
- They had worked in the same facility for at least 10 years, often sharing caseloads. They may have had a similar style of rating prior to the study.
- They each had ≥ 20 years clinical experience, and both had worked in the geriatric/continuing care field for the majority of that time.
- In sample 1 the EFAT-2 was administered regularly over a 4½ month time period. The raters became familiar with the tool by using it repeatedly during a week. This was consistent with one way the tool would be used in a clinical setting and had the advantage of allowing the raters to become comfortable with the tool. Both familiarity and regular use may positively impact reliability.

On the other hand, the formally trained raters had different amounts of clinical experience in different areas.

- One rater had used the EFAT-2 clinically on the acute PC unit for a 1 year period whereas the other rater was not familiar with the tool.
- Their work environments were different; one rater worked with the community-dwelling elderly while the other worked part-time in the public school system with disabled children and part-time on the acute PC unit.
- One rater had 20 years experience, the last 10 in geriatrics, and the other had 7 years experience, the most recent in PC.
- Due to personal commitments of the raters, the EFAT-2 was administered at irregular intervals over a 4½ month period in sample 2. This sporadic use of the tool had the potential to decrease the reliability of the raters.

The similarities between the raters of sample 1 and the dissimilarity between the raters of sample 2 are further illustrated statistically. Inspite of the high ICCs obtained by both pairs of raters there is an interesting difference in the level of significance of the F-test used for the repeated measures ANOVAs. The p value for the self-trained raters is not significant suggesting there is little difference between the total scores of the two raters. The p value for the formally trained raters is significantly different suggestive of systematic bias. Reviewing the contingency tables from sample 2, R3 scored subjects consistently higher than R4. On twelve occasions where disagreement was apparent the rating was greater than or equal to 2 rating points. One possible explanation for the disparity is the clinical background of the two raters. Clinicians tend to accommodate to the patient populations they work with. The paradigms - standards and expectations – align with the majority of the patients/clients the clinicians interact with. When they encounter a different population they require time to adjust their ideas as to what is

acceptable functional status and where it fits on the continuum of functional status as they know it. As a result, R3 and R4 may have used different criterion levels when rating patients on the EFAT-2.

Considering the problem that an undetected bias could cause may highlight the importance of acknowledging systematic bias. For example, if R3 and R4 used the EFAT-2 in a pre/post design study, where R3 saw subjects for the "pre" measurement and R4 saw subjects for the "post" measurement, the results might suggest that the individuals had improved. In fact, the "improvement" may be because of the consistent higher/lower scoring of the two raters. For this reason it is important that a systematic bias is taken into account.

Clinical experience, then, may affect the manner in which PTs perceive their patients and it may in turn affect their approach to patients. It appears that in the present study, disparity in scoring of some items may be due to the raters' having no previous knowledge of the subject, or having different clinical backgrounds. The disparity is particularly evident in items that are subjective.

In the focus group discussion, motivation was identified as an item all raters found difficult to measure. Motivation displayed low Kappa values in both samples but more so in sample 2. The low Kappa values may have been anticipated because motivation is a subjective attribute. Still, it is interesting to note the disparity between the pairs of raters. The self-trained raters had a comparatively low but acceptable Kappa value for motivation (0.53) while the formally trained raters had an extremely low Kappa value (0.17). Although the two samples differed in age (sample 2 was younger), they were similar regarding gender, diagnosis, and cognitive status. Assuming that

variation between samples is not a factor in decreasing reliability, one possible explanation of this difference would be that the circumstances of the raters had an effect on reliability.

The low Kappa values for the communication item are misleading and a possible explanation for this follows. Kappa values are affected by "prevalence" of the attribute to be measured.[38] For example in sample 1, communication ability was quite similar for all subjects. This resulted in a lack of variability in scoring. In fact both raters used only the 0 and 1 ratings. The lack of variability increased the likelihood of the raters agreeing just by chance. In sample 2, only categories 0, 1, and 2 were utilized. This too suggests diminished variability in scoring. Hence the Kappa statistic, corrected for chance agreement is low although percentage of agreement is high.

Inspite of the differences between the pairs of raters, both demonstrated excellent inter-rater reliability. This would suggest that therapists with a variety of clinical experience and different backgrounds might obtain acceptable ICCs when using the EFAT-2. Based on the two samples, the EFAT-2 can be a reliable instrument used on a regular basis or as a research tool where clinicians/research assistants are seeing subjects for the first time. The latter use might apply in some clinical settings. For example, in a rural setting a therapist may not work in the same facility every day of the week. The EFAT-2 could be used reliably on a weekly basis to monitor the ongoing functional status of patients.

5.2 Focus Group

Sim and Snell [45(p.189)] define a focus group as "a group interview – centered on a specific topic ('focus') and facilitated and coordinated by a moderator or facilitator –

which seeks to generate primarily qualitative data, by capitalizing on the interaction that occurs within a group setting." The objective of the focus group in this study was to determine whether the PTs who volunteered as raters considered the EFAT-2 a clinically useful tool. Although change will not be implemented solely on the strength of the opinions of the three raters, their comments will contribute to an evolving base of clinical evidence used in future development of the EFAT-2.

The items addressed by the raters in the focus group discussion: motivation, fatigue, and dyspnea, share a common attribute – they measure a subjective experience. In the clinical setting, the initial EFAT-2 would be administered once the patient had settled into the unit. Therefore the PT scoring the tool would have a general notion of the patient's motivation and fatigue level through chart notes and observation of the person on the unit prior to completing the EFAT-2. When the tool was used in an environment where the therapists worked, as in sample 1, the raters obtained higher reliability on individual items. For example, the self-trained raters appeared to be more consistent at rating fatigue (K= 0.916) than the formally trained (K= 0. 656). The self-trained raters had a sense of the subjects on the unit since both of them worked in the facility where the rating took place. R1 frequently referred to nursing notes to confirm her ratings and R2 worked on the continuing care PCU. The formally trained raters administered the EFAT-2 in facilities where they were not employed and based their scores on their first interaction with the subject and the long descriptors of the EFAT-2. The difficulty rating dyspnea experienced by R1 and R2 may have been due to a few factors. These may include: the ability of the raters to explain dyspnea rating, the subjects' ability to understand the concept of the dyspnea rating, or the higher

percentage of subjects with lung cancer. Dyspnea is a subjective experience and open to subjective interpretation.

It is well documented in the measurement literature that quantifying an "experience" is difficult. It was the opinion of the raters that motivation for example, was difficult to assess and provided little therapeutic information. Yet, it is an attribute all therapists must take into consideration and foster. Patients also attach considerable importance to aspects of their health other than those measurable by simple objective measurements.[45] Patients are concerned with the impact that health and illness have on their life, ie. with respect to their ability to contribute to family, or to continue their leisure pursuits etc. In health outcome measures, patient derived domains have included motivation and confidence.[44] Consideration of subjective attributes is advantageous when interacting with patients regardless of where they may be on the health continuum.

In any instrument or tool designed to categorize people there will be some individuals who do not absolutely fit the criteria for rating. The PS item, identified by the raters is a case in point. Often a subject would fit into one category but also met some criteria for another. When the PS item was used as a global rating based on the raters' overall impression of the subject as in sample 1, little difficulty was experienced in completing the item. Reliability was high and supports use of the global means of evaluation. The high Kappa value obtained when the raters' global impression of the subject was used suggests that clinical opinion may be used reliably in PC patients.

Clear guidelines are important for any clinical scale. The guidelines for the EFAT-2 appear well defined but were interpreted differently by each pair of raters. The

tool was designed to measure a specific performance – what an individual does on demand. Results of the focus group discussion suggest that the self-trained raters used the tool to reflect "usual" performance for the balance, mobility and locomotion items. For example, if a subject was able to transfer with one person assist during the testing (mobility rating "1") but they routinely had two health care providers assist with transfers on a day to day basis (mobility rating "2"), they were rated as a "2" by the raters. That is, R1 and R2 scored balance, mobility and locomotion according to the subjects usual performance.

Regardless of whether the EFAT-2 was used to rate usual performance or performance on demand, the Kappa statistic remained very high for the balance, mobility and locomotion items. These items are readily observable by the raters. The item behaviors and their underlying criteria were easy to describe in observable terms hence lessening subjective interpretation.[4] The EFAT-2 was designed to measure performance and should be used in that manner. If there is a discrepancy between the test performance and usual performance of a patient, the reason for the difference should be explored. The discrepancy is not a reflection of the measurement properties of the tool but may have relevance in a clinical setting. For example, if a patient regularly walks about the unit but on administering the EFAT-2 she/he refuses to, the reason for refusing should be explored. If there is a problem with symptom control, pyschosocial issues, or physical concerns they may be revealed and addressed.

5.3 Interpreting the tool clinically

One of the shortcomings of existing performance status measures used in PC is the tendency to cluster scores at the low end of the scale – the floor effect. This

produces a narrow range of scores and may not reflect differences in individuals in the PC environment. It was expected that the EFAT-2 would capture a greater range of scores for PC patients.

The range of total scores generated by the rater pairs (2-22 for R1, R2 and 3-27 for R3, R4), suggest that the EFAT-2 can capture a wider range of functional status due to the spread of total scores. We may be able to tell more about patients' functional status using the totals from the EFAT-2 than we could from the scores of the KPS or the ECOG. On the acute PCU the results of the EFAT are transcribed onto a graph (Appendix J). The visual representation facilitates interpretation of the tool. In essence, the darker the column, the more dependent the individual. The graphed EFAT is utilized as a teaching tool at some interdisciplinary team meetings on the acute PCU.

The purpose of the EFAT-2 is to evaluate functional status of PC patients. It has not been proven that the tool provides enough information on which to base treatment. Establishing a means of identifying the reason for a rating of 3 would be important if the tool were used to identify rehabilitation needs of PC patients. Further research is required into the EFAT's potential in this area.

5.4 Revisions to EFAT-2

The addendum to the guidelines will be incorporated into the Test Manual.

5.5 Directions for Future Research

The EFAT-2 is a reliable tool. It is necessary to ensure that it measures what it purports too. At this time it seems appropriate to look at the tool in terms of its validity. For example, how does the EFAT relate to other measures of disability. The Barthel index and the ERSS may be appropriate scales for establishing concurrent validity.

Reliability of the tool when used with a broader sample of patients on the PC continuum, i.e.: pain clinics and home care, could be examined. As well, reliability of the tool when used by other health professionals may bear investigation.

The tool's predictive validity could be investigated. Specifically, can the admission scores of the EFAT-2 be used to predict functional decline, length of stay or discharge location?

Difficulty in rating subjective items raises some questions. Should we be trying to "objectify" measurement of the items or look at other options for rating them, for example through the use of self report? It would be interesting to look at whether there is a correlation between results of "objective" ratings of the dyspnea, fatigue and motivation items and self report ratings of patients. Which method provides a more valid means of evaluating subjective experience?

One aspect of the EFAT-2 that has not been discussed is the instrument's ability to identify the rehabilitation needs of PC patients. Can the EFAT-2 be used to identify areas for rehabilitation intervention? When intervention is provided, can the EFAT-2 measure change? The test-retest reliability needs to be established for the EFAT-2 before further research in this area can be carried out.

Lastly, validation of the instrument in a broader context should be considered. Once the basic psychometric properties of the EFAT-2 have been established, the instrument should be tested in multi-center trials. Demonstrating the tools usefulness in different environments – other countries and in other languages will also enhance its utility.

CHAPTER 6
SUMMARY AND CONCLUSIONS

6.1 Summary

The EFAT was developed as a criterion - referenced outcome measure to evaluate the extent to which patients actually perform functional activities in a PC setting. The reliability and validity of the original version have been established.[4] A number of revisions were made to the EFAT in order to enhance its objectivity. Although the intent of the EFAT was not thought to have changed, the previously established reliability was no longer applicable. This study sought to establish reliability of the revised version, the EFAT-2.

The purpose of the study was to establish the inter-rater reliability of the EFAT-2 and to determine whether both formally trained and self-trained raters obtained an acceptable level of reliability. Two PTs who were trained in a similar way assessed PC patients on the EFAT-2. One pair of raters was self-trained, the other was formally trained. Two independent samples of 36 subjects were accrued.

Intraclass correlation coefficients were calculated to determine the inter-rater reliability between each pair of similarly trained raters. The 95% confidence interval and the SEM were also calculated for each pair of similarly trained raters. The Kappa statistic was used to determine the reliability of the rater pairs on individual items. A focus group was carried out to determine whether the physical therapists that administered the EFAT-2 considered it a clinically useful tool.

6.2 Conclusions

On the basis of the study conducted and presented above, the following conclusions were made:

1. The EFAT-2 has an excellent level of inter-rater reliability when used by PTs in the same manner as in this study.
2. Both self-trained and formally trained raters may obtain an acceptable level of reliability using the EFAT-2.
3. Non chance agreement on individual items for similarly trained raters, was acceptable for 9/10 items of the EFAT-2.
4. The EFAT-2 is easy to administer, quick to use and does not pose a burden to PC patients in an institutional setting.

REFERENCES

1. Tulman L, Fawcett J, McEvoy MD. Development of the Inventory of Functional Status - Cancer. *Cancer Nursing* 1991: 14: 254-60.
2. Hockley J. Rehabilitation in Palliative Care - Are we asking the impossible? *Palliative Medicine* 1993: 7(suppl): 9-15.
3. Hutchison TA, Boyd NF, Feinstein AR, Genda A, Hollemby D, Rowat B. Scientific Problems in Clinical Scales as Demonstrated in the Karnofsky Index of Performance Status. *J Chron Dis* 1979: 32: 661-6.
4. Kaasa T, Loomis J, Gillis K, Bruera E, Hanson J: The Edmonton Functional Assessment Tool: Preliminary development and evaluation for use in palliative care. *J Pain and Symptom Manage* 1997: 12: 1-10.
5. Portney LG, Watkins MP. *Foundations of Clinical Research Applications to Clinical Practice*, East Norwalk, Connecticut: Appleton and Lange 1993.
6. McDowell I, Newell C. *Measuring Health. A guide to rating scales and questionnaires*. 2nd ed. Oxford University Press 1996.
7. Rothstein JM: Measurement and clinical practice : Theory and application. IN: Rothstein JM, ed. *Measurement in Physical Therapy* New York : Churchill Livingston. 1985: 1-46.
8. Connelly DM, Stevenson TJ, Vandervoort AA. Between and within-rater reliability of walking tests in a frail elderly population. *Physiother Can* 1996: 48: 47-51.
9. Fulton CL. Physiotherapists in cancer care. A framework for rehabilitation of patients. *Physiother* 1994: 80: 830-834.
10. Thorpe G: Introduction to Rehabilitation. IN: Doyle D, Hanks GWC, MacDonald N, eds. *Oxford Textbook of Palliative Medicine*. Oxford: Oxford University Press. 1993: 529
11. Fulton CL. Measures of rehabilitation status in oncology. *Physiother* 1994: 80: 849-853.

12. Karnofsky DA, Abelman WH, Craver LF, Burchenal JH: The use of nitrogen mustard in the palliative treatment of carcinoma. *Cancer* 1948: 1: 634-656.

13. MacDonald N. Principles governing the use of cancer chemotherapy in palliative medicine. IN: Doyle D, Hanks GWC, MacDonald N, eds. *Oxford Textbook of Palliative Medicine*. Oxford: Oxford University Press. 1993: 105-117.

14. Anderson F, Downing M, Hill J, Casorso L, Lerch N. Palliative Performance Scale (PPS): A new tool. *J Palliat Care* 1996: 12(1): 5-11.

15. Yates JW, Chalmer B, McKegney PF. Karnofsky performance status. *Cancer* 1980: 45: 2220-2224.

16. Evans C, McCarthy M. Prognostic uncertainty in terminal care: can Karnofsky index help? *Lancet* 1985:1: 1204-1206.

17. Croos V, Waller S, Smith T, Hahn TJ. The use of the Karnofsky Scale in determining outcomes and risk in geriatric outpatients. *J Gerontol* 1991: 46: M139-144.

18. Coscarelli Schag C, Heinrich RL, Ganz P. Karnofsky performance status revisited: reliability, validity and guidelines. *J Clin Oncol* 1984: 2: 187-193

19. Soresen JB, Klee M, Hansen HH. Performance status assessment in cancer patients. An inter-observer variability study. *Br J Cancer* 1993: 67: 773-775.

20. Cole B, Finch E, Gowland C, Mayo N. *Physical Rehabilitation Outcome Measures*. Toronto, Ontario: Canadian Physiotherapy Association. 1995: 54-55.

21. Fucile, J. Functional rehabilitation in cancer care. *Sem Oncol Nurs* 1992: 8: 186-189.

22. Marciniak CM, Sliwa JA, Spill G, Heinemann AW, Semik PE. Functional outcome following rehabilitation of the cancer patient. *Arch Phys Med Rehabil* 1996: 77: 54-57.

23. Schag CAC, Heinrich RL. Development of a comprehensive quality of life measurement tool: CARES. *Oncology* 1990: 4 (5): 135-138.

24. Schag CAC, Ganz PA, Heinrich RL. Cancer Rehabilitation Evaluation System - Short Form (CARES-SF). *Cancer*1991: 68: 1406-1413.

25. Affleck JW, Aitken RCB, Hunter JAA, McGuire RJ, Roy CW: Rehabilitation status: a measure of medicosocial dysfunction. *Lancet* 1988: 1: 230-233.

26. Maguire P, Selby P. Assessing quality of life in cancer patients. *Br J Cancer* 1989: 60: 437-440.

27. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica* 1983: 67: 361-370.

28. Mattison PG, Aitken RCB, Prescott RJ. Rehabilitation status – the relationship between the Edinburgh rehabilitation status scale (ERSS), Barthel Index, and PULSES profile. *Int Disabil Studies* 1991: 13: 9-11.

29. Mahoney F, Barthel DW. Functional assessment: the Barthel index. *Maryland State Med J* 1965: 14: 61-65.

30. Gresham GE, Labi MLC. Functional assessment instruments currently available for documentary outcome in rehabilitation medicine. IN: Granger CV, Gresham GF, eds. *Functional Assessment in Rehabilitation*. Baltimore: Williams and Wilkins. 1984: 65-85.

31. Cella DF, Tulsky DS. Measuring quality of life today: methodological aspects. *Oncology* 1990: 4: 29-38.

32. Badley EM, Lee J. Impairment, disability and the ICIDH model III:underlying disease, impairment and disability. *Int Rehabil Med* 1986: 8:174-181.

33. World Health Organization (WHO). *International classification of impairments, disabilities and handicaps: a manual of classification relating to the consequences of disease*. Geneva, Switzerland: World Health Organization. 1980.

34. Unpublished: Kaasa T, Wessel J, Turner K. The Edmonton functional assessment tool: Further development and validation for use in palliative care. Poster presentation at Research symposium in rehabilitation and related health services, September 1997 at the University of Alberta: Edmonton, Alberta

35. Streiner DL, Norman GR. *Health Measurement Scales. A practical guide to their development and use.* 2nd ed. Oxford: Oxford University Press 1995.

36. Dumholdt E. *Physical Therapy Research. Principles and Applications.* Philidelphia: WB Saunders Co. Harcourt Brace Jankovich Inc. 1993

37. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psycho Bull* 1979: 86: 420-428.

38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977: 33: 159-174.

39. Personal Communication. Brenneis C, Program Director Edmonton Regional Palliative Care Program, MacKay S, Palliative consult nurse, April 15, 1997.

40. Stratford PW. Confidence limits for your ICC. Phys Ther 1989: 69: 237-238

41. Denegar CR, Ball DW. Assessing reliability and precision of measurement: An introduction to intraclass correlation and standard error of measurement. *J Sport Rehabil* 1993: 2: 35-42

42. Bruera E. Ethical issues in palliative care research. *J Palliat Care* 1994: 10(3): 7-9.

43. Kristjanson LJ, Hanson EJ, Balneaves L. Research in palliative care populations: Ethical issues. *J Palliat Care* 1994: 10(3): 10-15.

44. MacDonald N. Quality of life in clinical and research palliative medicine. *J Palliat Care* 1992: 8 (3): 46-51.

45. Sim J, Snell J. Focus groups in physiotherapy evaluation and research. *Physiother* 1996: 82: 189-198.

APPENDIX A
EDMONTON FUNCTIONAL ASSESSMENT TOOL

EDMONTON FUNCTIONAL ASSESSMENT TOOL (EFAT-2)

	0 Functional	1 Min Dysfunction	2 Mod Dysfunction	3 Severe Dysfunction	D a t e			
Communication	Independent with all aspects of communication	Communicates effectively > 50% but < 100% of time	Communicates effectively <50% of time	Unable to communicate				
Mental Status *	Oriented x 3 Memory intact	Impair 2/6 orientation/memory. Follow simple commands	Impair 3-4/6 orientation/memory. Responds inconsistently or restless, agitation, anxious	Impair 5-6/6 orientation/memory or unresponsive to verbal commands				
Pain	None or occ. pain. Pain does not impact function'	Pain limits some activity. Inhibits function minimally	Pain present all the time. Inhibits function mod.	Unable to do any activities because of pain				
Dyspnoea *	No dysfunction	Urgency = counting or SOBOE or intermittent	1 extra breath with counting or O_2 at 1-3 litres	≥ 2 breaths with counting or O_2 at ≥ 4 litres				
Balance * Sit Stand	Independent Balance	↓ balance. Attain/maintain position with equip or 1 person. Min. safety risk	Unsafe balance. Maintain position with mod. assist 1 or more. Risk of fall	Maintain position with max assist 1-2 persons or unable to evaluate				
Mobility *	Controls/moves all limbs at will. Performs safely and independently	Control/move all limbs but degree of limitation. 1 assist to move/safety	Can assist another person who initiates movement. Requires 2 persons assist for safe transfer	Unable to assist with position change. Mechanical lift to transfer				
Locomotion ** Walk Wheelchair	Walks unassisted or independently in lead up and propelling	Walks with 1 person assist/± walk aid or supervision with lead up	Walks with 2 person assist short distance or requires assist with lead up/propel wheelchair	Unable to walk. WB for transfer. Dependent W/C management				
Fatigue ***	Rarely needs to rest	Rest <50% of day	Rest >50% of day	Bedridden due to fatigue				
Motivation ***	Wants to participate despite limitations	Active/passive participant >50% of time	Active/passive participant <50% of time	No desire to participate in activity				
ADL	Independent	Independent using adaptive equipment	Manual assist of 1, verbal cueing/supervision to complete task	Total assist with ADL				
Performance Status	Independent in room or unit	Independent with minimal assist of 1	Mod assist of 1 person room/unit	Assist of 1-2 persons in room				

Sensation - deleted * - reworded to ↑ objectivity ** - Activity/WC mobility collapsed *** - New items

APPENDIX B

EDMONTON FUNCTIONAL ASSESSMENT TOOL (EFAT)

FOR PALLIATIVE CARE

TEST MANUAL

Terry Kaasa, B.P.T.

Kathy Gillis, BSc.OT.

1993

The authors acknowledge Joan Loomis, Associate Professor in the Department of Physical Therapy at the University of Alberta, for her assistance with analyzing the data and writing the test manual while she served as a Research Therapist in the Rehabilitation Medicine Department at the Edmonton General Hospital from September to December, 1993.

The authors are grateful to Barbara Iwaniuk, BSc.OT., for serving as the second rater in the reliability study.

TEST MANUAL
EDMONTON FUNCTIONAL ASSESSMENT TOOL (EFAT)

NEED

The purpose of palliative care (PC) is to improve, not to cure. The majority of clients with terminal cancer have an impaired ability to perform everyday functions at different stages of their disease. Often when initial symptoms abate, clients' ability to function will improve for a time.

Even though a change in functioning is seen through the course of PC, it is seldom described in the literature (Hutchinson et al., 1974). Yet, clients seek medical advice due to difficulties in performing everyday functional activities. A measure of functional status/performance is needed for a more comprehensive evaluation process with clients with terminal cancer. Furthermore, a measure of functional status/performance is needed to demonstrate the impact that rehabilitation and medical interventions have on functional status.

A systematic evaluation of functional performance is important in the palliative care setting to provide objective information to identify the individuals' capabilities to enhance quality of life and to compare functional status over time.

BACKGROUND

In 1990 the rehabilitation staff in the palliative care unit at the Edmonton General Hospital recognized a need for an evaluation of physical function of clients with terminal cancer. The existing database was not meeting the needs of the clients because it tended to focus on identification of impairments and prompted treatment planning from a curative approach. Moreover, frequent reassessment required by this group and comparison of functional performance over time was not possible. Nor did it satisfactorily communicate client status to the interdisciplinary team.

The rehabilitation staff began to search for an appropriate evaluation tool following guidelines by Cella and Tulsky (1990). They recommended that a measure should contribute significantly to patient care, not pose burden to patients and staff, be responsive to physical and occupational therapy functions, and be perceived as unobtrusive by the patients and families. Additionally, the staff wanted an evaluation which would focus on the clients' abilities, allow repeated assessments, and communicate concise and relevant information to the team.

A search of the literature was undertaken. Rubenstein et al. (1989) defined functional status as the end result of a person's health (absence of disease), well-being (capacity to participate in life) and coping (capacity to overcome health problems). Functional performance was defined as a multidimensional concept that encompasses personal care, general activity level and cognition (clear thinking, attention, memory).

A literature review of evaluation of functional performance in oncology and PC revealed two widely used scales. Karnofsky Performance Status (KPS) and Eastern Cooperative Oncology Group (ECOG) were considered "gold standards" for assessment of performance status. Both scales have demonstrated reliability and validity.

KPS was developed by Karnofsky and Burchenal in 1949 to measure a client's productivity and dependence on custodial or medical care. Grieco and Long (1984) and Connil et al. (1990) have suggested that the KPS was a reliable global measure of performance. Mor et al. (1984) determined interrater reliability at .97 for trained observers. Yates et al. (1980) showed construct and predictive validity for persons with cancer. Hutchinson et al. (1974) pointed out that the KPS lacked operationally defined items and scale points. Our own study revealed that KPS score tended to cluster at the low end, limiting sensitivity to monitor change. Further, the KPS did not meet the needs of rehabilitation assessment because it did not evaluate functions addressed by rehabilitation.

Functional evaluations in the rehabilitation literature were reviewed (Granger, 1979). The Barthel and Katz Indexes, even though designed for clients with chronic disease, were unsuitable for our purposes in PC because only personal care was assessed.

Quality of life measures (QOL) were also reviewed. The multidimensional nature of QOL measures were too broad in scope to meet the need for a concise tool to specifically measure functional status. McCartney and Larson (1987) found that QOL scores in persons with gynaecological cancer tended to become quite low by nine weeks before death, leaving little room to reflect change.

An appropriate evaluation tool was not found. Subsequently, the rehabilitation members of the palliative care team developed an instrument to meet their needs based on observations of functional performance of clients in palliative care at the Edmonton General Hospital.

PURPOSE

The Edmonton Functional Assessment Tool (EFAT) was designed to evaluate functional performance of clients with terminal cancer. The purpose of the EFAT is to evaluate the degrees of functional performance of clients throughout the terminal phase no matter how limited. The EFAT assesses functional performance of a single client over time and allows comparisons between groups of clients.

DEVELOPMENT OF THE EFAT

The EFAT consisted of ten functions that the physical therapist and occupational therapist on the palliative care unit at the Edmonton General Hospital determined to be important to clients' physical functioning. A minimum number of functions were included in the EFAT for ease of administration, yet reflect the range of functional performance still seen in this population even in the final stage. Each item was defined in terms of the behaviors involved in performing each function

1. communication- is the ability to speak, write and/or read; to use alternate communication devices, as required (eg. glasses, hearing aid, vice amplifier communication board); to use a telephone with or without adaptation

2. mental status- involves identifying self, place and time; recalling immediate, recent and remote events; and responding to questions or following instructions accurately

3. pain- is the extent to which physical suffering limits or interferes with the clients' ability to participate in desired activity

4. sensation- pertains to the ability to identify and localize light touch, pressure and temperature

5. respiratory status- is the degree to which dyspnoea limits the client's ability to participate in daily activities

6. balance- is the ability to maintain or regain a position of sitting or standing against slow or fast displacement

7. mobility- is the ability to roll, bridge, attain sitting, and transfer from bed to chair or commode or toilet

8. activity level- is the client's ability to ambulate

9. wheelchair mobility- is the ability to manage wheelchair parts (armrests, footrests, brakes and seat belt); to propel the wheelchair on and off the unit, to open and close doors and wheel through doorways

10. activities of daily living- encompasses the ability to wash, dress, comb hair, apply make-up, brush teeth, shave, toilet and feed one's self

Each item in the EFAT was evaluated by a 4-point rating scale from 0 to 3. Each rating described the behaviors that a client must demonstrate to achieve each rating. These descriptors were gradations of functioning with 0 representing functional independent performance and 3 representing total loss of functional performance. Given the ultimate physical decline of this population, the 3 rating was used when an item could not be assessed or it was unsafe to do so. A total possible score on the EFAT was 30.

In addition to the EFAT, a global performance status rating (PS) was devised. It asked for an overall judgment of functional performance taking into account the ten functions assessed by the EFAT.

TESTING THE EFAT

Testing the EFAT followed three steps
1. pilot testing
2. validation study
3. revision of the EFAT and ongoing research

1. Pilot testing

The EFAT was tested on 39 clients, who were admitted consecutively to the palliative care unit at the Edmonton General Hospital, to ascertain the feasibility of the instrument, administration procedures and instructions prior to the validation studies. The clients were assessed on admission by either the physical or occupational therapist using the EFAT and reassessed once a week until transfer off the unit, discharge or death occurred. The usual team assessments continued and included the Edmonton Symptom Assessment Sheet (ESAS) and Mini-Mental State Exam (MMSE), KPS and ECOG. The results were promising. The pain and respiratory items had ambiguous descriptors and were revised. The EFAT seemed to add new information, yet it was related to existing evaluations. The total EFAT score correlated with the KPS (r= - 56) and ECOG (r=.69). The EFAT mental alertness score correlated minimally with the MMSE score (r=.317) and moderately with the ESAS drowsiness score (r=.571). The EFAT pain score correlated minimally (r=.37) with the pain visual analogue scale. The EFAT activity score correlated moderately with an activity visual analogue score (r=.47).

2. EFAT validation study

The EFAT was tested on 101 clients admitted consecutively to the palliative care unit at the Edmonton General Hospital from September 1992 to September 1993. In addition to the usual assessments conducted by the team, the clients were assessed within three days of admission by either the physical or occupational therapist using the EFAT. Reassessments were done by either therapist once a week until transfer off the unit, discharge or death occurred. Clients who died or were transferred formed the unit group (n=88), and clients who were discharged formed the home group (n=13). These two groups were analyzed separately to estimate construct validity. It was predicted that the unit group will have higher admission EFAT scores than the home group, that the EFAT scores will increase over the course of the disease from the initial assessment to the final assessment, and that the difference in EFAT scores from initial to final assessment will be greater of the unit group than the home group.

A subgroup of this sample, admitted from June to September 1993, were assessed by two different therapists using the EFAT, EFAT global performance status (PS), KPS and

ECOG. This subgroup contained 25 clients. The same two raters evaluated each client separately within 24 hours at 3-4 days after admission. The purpose of this subgroup was to estimate the inter-rater reliability and concurrent validity of the evaluations.

Description of the sample

The sample of 101 clients consisted on 43.6% females and 56.4% males with an average age of 63.9 years.The location of cancer varied with the 79.1% of the cancer occurring in the gastrointestinal tract (27.7%), lung (22.7%), genitourinary tract (17.8%), and breast (10.9%). Ninety-one of these clients were admitted to acute palliative care and 10 clients were admitted to long-term palliative care. This sample was later classified into the unit group (n=88) and the home group (n=13). The average age of the unit group was older at 65.6 years compared to 62.1 years for the home group, but this difference was not significant (p=.09). The frequency of cancer type was comparable for the two groups.The average length of stay was longer at 48.1 days for the unit group compared to 28.5 days for the home group, and this difference was also not significant (p=.13). The frequency of sex was vastly different for the home group consisting of 12 males and 1 female. This figure may suggest that discharge home depended on availability of a caregiver.

The subgroup of 25 clients in the unit group had a similar frequency of sex and location of cancer as the larger group. The average age of 67.1 years was slightly higher than the larger group.

Reliability of the EFAT

The inter-rater reliability of the total EFAT expressed as an intraclass correlation (ICC) was .883 (p=.0026). Even though the EFAT was moderately reliable, the mean total score given by rater 1 ($\bar{x}$=37.24) was significantly different from the mean total score given by rater 2 ($\bar{x}$=42.92) (p=.05). An inter-rater reliability of .883 is comparable to other evaluations of performance status used in oncology and palliative care (Mor et. al. 1987). By comparison in our sample, the intraclass correlations between the same two raters for the KPS was .71 and for the ECOG was .81.

The inter-rater reliability coefficients (ICC) for seven of the ten items of the EFAT ranged from .70 to .93. Items with lower ICC values were sensation at .20, pain at .52 and mental alertness at .61.

Validity of the EFAT

Content validity was safeguarded by collaborative planning between the physical therapist and occupational therapist during the development phase. The ten items appeared to represent the difficulties in functional performance of the clients seen in palliative care.

Concurrent validity of the EFAT was demonstrated by correlating the total EFAT score with other performance measures used in the field: KPS and ECOG. The Pearson Product

Moment correlations at -0.79 (p=.0001) and .85 (p=.0001), respectively, suggested that the total EFAT score was associated with the types of performances included in the KPS and ECOG. The total EFAT score was also strongly related to its accompanying global PS rating at r=.90 (p=.0001).

Construct validity of the EFAT was supported by confirmation of the predictions. The final mean EFAT score was significantly higher than the initial mean EFAT score for the total sample of 101 clients (t=-3.216, p=.01). This finding indicated a deterioration in function over the progression of the disease.

The difference in EFAT scores from initial to final assessment was significantly greater for the unit group (n=53), [t = -7.509, p=.0001], but this difference was not significant for the home group (n=7), [t=.072, p=.94].

Clients in the home group showed significantly lower mean admission scores on the EFAT ($\bar{x}$ = 26.46) than clients in the unit group ($\bar{x}$ = 40.78) (t test for unequal samples =3.25, p=.0046). Seven of the ten EFAT items were significantly different in the predicted direction. The items that did not follow the predicted direction were pain, sensation and wheelchair mobility.

Taken together, these findings suggest that the EFAT measures what it purports to measure.

3. Revision of the EFAT and ongoing research (second edition)

Revisions were made to the EFAT based on the findings of the validation study. Descriptors for respiratory function, balance, mobility and mental alertness were reworded to clarify behaviors to be observed. The sensation item was deleted because it was an impairment that did not relate to the total EFAT score. Even though the pain item had poor reliability and did not relate to the total EFAT score, it was retained because of its perceived value to quality of life and validity of an evaluation used with the palliative care population. The items on activity level and wheelchair mobility were collapsed into one item called locomotion, mainly because the wheelchair mobility item was seldom used. The new locomotion item can be used if the client is either mobile on foot or in a wheelchair. Two items on fatigue and motivation were devised.

DESCRIPTION OF THE SECOND EDITION OF THE EFAT

The second edition of the EFAT consists of ten functions. Each item and its four levels of performance (0-3) are defined in behavioral terms as follows:

1. **communication:** is the ability to speak, write and/or read; to use alternate communication devices, as required (eg. glasses, hearing aid, vice amplifier, communication board); to use a telephone with or without adaptation.
   - **0** functional : is independent with all aspects of communication including reading, writing and verbal expression.
   - **1** minimal dysfunction: able to communicate effectively greater than 50% of the time but less than 100% of the time due to confusion, dysarthria, weakness, fatigue, dyspnea or language barrier.
   - **3** moderate dysfunction able to communicate effectively less than 50% of the time due to confusion, dysarthria, aphasia or weakness
   - **3** severe dysfunction unable to communicate with others due to severe weakness or unresponsiveness

2. **pain** is the extent to which physical suffering limits or interferes with the clients' ability to participate in desired activity.
   - **0** functional: has no or occasional pain which does not impact on function
   - **1** minimal dysfunction: may have "bad days" when pain will limit some activity and inhibit function minimally
   - **2** moderate dysfunction: has pain the majority of the time and inhibits function moderately
   - **3** severe dysfunction has severe continuous pain and is unable to do any activity or function

3. **mental status** is the ability to identify (1) self, (2) place and (3) time; to recall (4) immediate, (5) recent and (6) remote events; to respond to questions and to follow instructions accurately.
   - **0** cognitively intact, oriented, has intact memory AND responds to all stimuli and understands explanations
   - **1** minimal dysfunction impaired in 2 out of 6 orientation and memory tests AND able to follow simple commands
   - **2** moderate dysfunction impaired in 3-4 out of 6 orientation and memory tests AND responds to commands inconsistently OR is restless, agitated, aggressive or anxious
   - **3** severe dysfunction impaired in 5-6 orientation and memory tests OR is unresponsive to verbal commands

4. **dyspnoea** is the degree to which shortness of breath limits the client's ability to participate in desired activity.
   - **0** functional does not impact on participation in daily activity AND counts to 15 at a rate of 2 counts per second AND does not require oxygen

1 minimal dysfunction has urgency to breathe after counting consecutively to 15 OR has increased use of accessory muscles of breathing, AND has shortness of breath on exertion OR requires intermittent use of oxygen
2 moderate dysfunction requires one additional breath to complete counting to 15 OR uses oxygen continuously at 1-3 litres
3 severe dysfunction requires 2 or more breathes to complete counting to 15 OR cannot speak at all due to dyspnoea OR uses oxygen continuously at greater than 4 litres

5. **sitting or standing balance**is the ability to maintain or regain the position of sitting or standing against slow or fast displacement.
   0 functional: independent, safe balance
   1 minimal dysfunction: has decreased balance AND is able to maintain the position with use of equipment or assistance of one person AND is at minimal safety risk
   2 moderate dysfunction: has unsafe balance AND requires moderate assistance of one or two persons to maintain the position AND is at risk of falling
   3 severe dysfunction: requires maximum assistance of one or two persons to maintain the position OR unable to evaluate balance due to safety concerns or severe systemic illness

6. **mobility** is the ability to move and control the body for rolling, bridging, attaining sitting from side lying and transfers from bed to chair, commode or toilet.
   0 functional: with or without aids is able to control and move all limbs at will AND turns, lifts, balances and attains sitting and standing ad lib. AND may use side rails or monkey bars AND performs tasks safely and independently
   1 minimal dysfunction: is able to move and control all limbs but a degree of limitation in range of motion or strength exists AND requires assistance of one person to turn , to attain sitting or standing and to balance AND initiates requests for help AND performs tasks safely with assistance of one person
   2 moderate dysfunction: can assist another person who initiates movement by turning, lifting, balancing and/or sitting up AND requires assistance of two persons to transfer safely
   3 severe dysfunction: does not assist in any way to change position AND is completely dependent on others for movement AND requires a mechanical lift for transfers

7. **walk or wheelchair locomotion**is the ability to walk OR is the ability to manage wheelchair parts (eg. armrests, footrests, brakes, seatbelt) and to propel the wheelchair safely on the unit.
   0 functional: walks safely without human assistance, may use cane or walker, AND rises from bed or chair on own OR is able to perform all lead-up activities AND can propel wheelchair independently and safely on the unit
   1 minimal dysfunction: walks safely with assistance of one person AND may uses cane or walker AND rises from bed or chair on own OR requires supervision with lead-up activities but propels chair independently and safely on the unit

2 moderate dysfunction: walks safely with assistance of two persons for short distances OR walks to chair at bedside only with assistance of two people OR requires assistance with lead-up activities and to propel wheelchair
3 severe dysfunction: is unable to walk but may weight bear for transfers OR is unable to perform any aspects of wheelchair management

8. **activities of daily living (ADL)** is the ability to wash, dress, comb hair, apply makeup, brush teeth, shave, toilet and feed oneself.
   0 functional: is independent with all aspects of ADL
   1 minimal dysfunction: is independent with all aspects of ADL using adaptive devices (eg. reacher, shoe horn, walker, raised toilet seat, commode, eating utensils etc.) OR requires occasional verbally cues to complete tasks
   2 moderate dysfunction: requires intermittent manual assistance AND/OR constant verbal assistance to complete tasks
   3 severe dysfunction: requires total assistance with all ADL

9. **fatigue** is the subjective feeling of weariness, weakness, exhaustion, lack of energy resulting from prolonged stress directly or indirectly due to the disease process.
   0 feels tired or fatigued OR feels tired periodically OR rarely needs to rest
   1 minimal dysfunction: feels quite tired and fatigued on occasion OR needs to rest less than 50% of the normal daytime
   2 moderate dysfunction: usually feels very tired and fatigued OR needs to rest greater than 50% of the normal daytime
   3 severe dysfunction: feels exhausted most of the time OR is unable to get out of bed due to fatigue

10. **motivation** is the ability to initiate movement or behavior, induces a person to act, desire
   0 functional: demonstrates initiative with daily activities AND is actively involved in decision-making, goal setting and treatments AND wants to participate despite limitations
   1 minimal dysfunction: requires minimal encouragement to participate in activities requires minimal assistance with decision-making, goal setting and treatments ANDactively/passively participates greater than 50% of the time (minimal loss of interest or pleasure in usual activities)
   2 moderate dysfunction: requires moderate encouragement to participate in activities AND requires moderate assistance with decision-making, goal setting and treatments AND actively/passively participates less than 50% of the time (moderate loss of interest or pleasure in usual activities)
   3 severe dysfunction: shows no or minimal compliance with assessments and treatments AND requires maximal assistance with decision making and goal setting AND does not express desire to participate in activities.

Performance status is a global rating of the ten functions assessed by the EFAT. The four descriptors for performance status are

0 -is independent in the room or unit, and may also be independent off the unit
-ambulates inside and/or outside of the room ad lib. (may or may not use equipment, eg. wheelchair, walker)
-is independent with activities of daily living, may wear own clothes
-pursues leisure activities independently, demonstrates initiative and motivation

1 -is independent with minimal assistance of one person in the room
-attends activities off the unit with assistance of a person
-walks or uses wheelchair for short distances with or without supervision and at occasions throughout the day
-requires assistance of one person to balance safely in sitting or standing
-requires assistance of one person to move sit up and stand up.
-is independent with activities of daily living using adaptive devices
-partakes in leisure interests through structured activity on the unit
-spends less than 50% of the time in bed

2 -requires moderate assistance of one person in the room or unit
-can slightly change position of the body or extremities but is unable to make frequent or significant changes independently
-walks a few steps or not at all
-unable to bear weight on legs and/or must be assisted into a chair or wheelchair
-is dependent with wheelchair setup but propels chair short distances independently
-carries out modified activities of daily living with the assistance of one person
-requires assistance of one person to partake in leisure activities
-spends greater than 50% of the time in bed

3 -requires assistance of 1-2 persons in the room
-confined to bed and unable to change position
-is totally dependent with activities of daily living
-rarely observes activities on the unit and pursues passive forms of leisure, eg. listening to tapes, being read to

INSTRUCTIONS FOR USE OF THE EFAT

1. The EFAT is an indicator of functional performance with clients who have terminal cancer in a hospital palliative setting. Caution is advised if the EFAT is used in other palliative care settings.

2. Ask the client to do each function (item) in turn and directly observe performance. Choose the rating that best describes the client's performance when tested, even if it is at odds with indirect observations on the unit. A discrepancy in client performance indicates a need for the interdisciplinary team members to discuss reasons for the client's usual behavior on the unit. The EFAT measures what the client <u>can do on request</u>, not what he or she ought to be able to do.

3. Rate the client's performance only on the specific functions defined by each item. For example, wheelchair locomotion includes managing the wheelchair parts and propulsion, it does not include a wheelchair transfer. Transfers are included under mobility. In item 5 indicate whether sitting or standing balance is tested. In item 7 indicate whether walking or wheelchair locomotion is tested.

4. Read the definitions of each item carefully when scoring the EFAT. Choose one rating of performance, from the four choices (0-3) under each item, which best describes the client's performance. If the descriptor lists tasks connected by the word AND, the client must complete all tasks. If the descriptor lists tasks connected by the word OR, the client must complete one task. If the descriptor lists tasks connected by both AND and OR, the client must complete some of the tasks.
   For example, the descriptor for locomotion with a wheelchair requires the client to perform both lead-up activities AND propulsion of the wheelchair without assistance of a person in order to obtain a '0' rating. The descriptor for ADL describes that the client may require either manual assistance OR verbal assistance to complete the task to receive a '2' rating.

5. Evaluate all items. The total score on the EFAT is 30. No provision is made to consider an item 'not applicable' or 'not evaluated'. The '3' rating is used if you are unable to assess or if the client would be at risk of harm.

6. Complete a rehabilitation database as required to plan appropriate intervention. The EFAT is functional outcome measure, hence does not include all domains that can be potentially measured for prescriptive purposes.

LIMITATIONS OF THE EFAT

1. Raters involved in the validation study have tended to be the developers of the tool.
2. The EFAT should only be used with the palliative population for which it was developed and tested.
3. Some items remain subjective despite attempts to define behaviors impacting on functional performance, eg. pain.
4. Use of the total score is limited for prescribing treatment because two clients with the same total EFAT score may present differently in terms of performance on individual items.
5. Guidelines for scoring the EFAT were not documented prior to this test manual.

ADVANTAGES OF THE EFAT

1. The EFAT item scores and the total EFAT score are sensitive to change in functional performance and can be used to monitor this change over time.
2. The EFAT can be administered quickly without tiring clients due to inclusion of a minimum number of functions in the test.
3. The EFAT can be easily administered by raters from any health discipline after carefully studying the test guidelines.
4. Reliability and validity coefficients are established for the first edition of the EFAT.

CONCLUSIONS

The EFAT was developed as an outcome measure to evaluate effectiveness of palliative care interventions. The EFAT has shown to be a reliable and valid measure of functional performance in the palliative care population. The EFAT has been tested sufficiently and should be used in controlled clinical trials designed to determine the functional outcomes of interventions in this population. Ongoing research should also be conducted to improve the measurement properties of the tool.

The developers of the EFAT would like to receive your feedback and suggestions to assist with the ongoing development of the tool. Please forward your suggestions for improvement or any problems with collecting and recording data to Terry Kaasa, Physical Therapist, Department of Rehabilitation Medicine, Cross Cancer Institute, Edmonton, Alberta.

APPENDIX C
REGIONAL PALLIATIVE CARE PROGRAM
CRITERIA FOR ADMISSION

Palliative Care Is:

Active total care offered to a patient with progressive disease and their family when it is recognized that the illness is no longer curable, in order to concentrate on the quality of life and the alleviation of distressing symptoms in the framework of a coordinated service. Palliative Care neither hastens nor prolongs death. It provides relief from pain and other distressing symptoms and integrates the psychological and spiritual aspects of care. In addition, it offers a support system to help relatives and friends cope during the patient's illness and bereavement.

(Medical/Nursing/Midwifery Advisory Committee – United Kingdom)

Based upon this definition, all people admitted to the program will:

- be experiencing progressive disease where the focus of care is on comfort, not cure, and improving their quality of life
- require active care to alleviate distressing symptoms related to physical, psychosocial and spiritual needs

Approximately 85-90% of these people will have a cancer diagnosis.

Admission Criteria to Specific Areas:

Home
- above criteria
- expected length of stay on the program of approximately 3-4 months
- do not require acute/tertiary care
- the ability to provide services within financial resources
- desire for the person/family to be cared for at home

Continuing Care
- above criteria
- cannot be managed at home
- do not require acute/tertiary care
- expected length of stay approximately 2 months
- over 18 years
- accepting of no code status

Acute Care
- for management of acute medical problems (ie. pathological fracture, bleed, acute respiratory distress)
- anticipated short stay

Tertiary Palliative Care Unit
- severe symptom problems for which management has not been successful in any of the other settings, and requiring intensive management
- expected length of stay of approximately 2 weeks
- over 18 years
- accepting of no code status

APPENDIX D
MINI MENTAL STATE EXAM/LOWER QUARTILES FOR REFERENCE

Capital Health
CHA-205-PC Dec 97

MINI-MENTAL STATE
(FORM A)

MAXIMUM SCORE	SCORE	
5		What is the (year) (season) (day) (month) (date)?
5		Where are we: (province) (country) (town) (hospital) (floor)?
3		Name three objects: glass blanket pencil
5		Serial 7's. Alternatively spell "World" backwards.
3		Ask for the 3 objects repeated above.
2		Name a pencil and watch.
1		Repeat the following - "No ifs, ands or buts".
3		Follow a 3-stage command: "Take a paper in your right hand, fold it in half, and give it to me".
1		Read and obey the following: **CLOSE YOUR EYES**
1		Write a sentence.
1		Copy a design.
30		**TOTAL SCORE**

Assess level of consciousness along a continuum.

Alert Drowsy Stupor Coma

Number of years of schooling: ?_____

Lower Quartiles	Age					
Schooling	≤39	40-49	50-59	60-69	70-79	≥80
0-4 years	20	20	20	19	18	16
5-8 years	24	24	25	24	23	22
9-12 years	28	28	27	27	26	23
College experience or higher degree	29	29	28	28	27	26

Date: _____

Reference : Crum RM, Anthony JC, Basset SS, Folstein MF. Population-based norms for the mini mental state examination by age and educational level. *J Am Med Assoc* 1993: 269: 2386-2391

APPENDIX E
SAMPLE SIZE CALCULATION

Sample Size Calculation†

$$n = v+1$$
$$= (p-po) / (1-p\ po)$$

where " p " is an estimate of the correlation we feel it is important to find

" po" is the value specified as a minimal acceptable correlation*

Subjects required to reject Ho: po = 0.8 at the 5% level of significance, 80% power:

$$= (p-po) / (1-ppo)$$

$$= (.90 - .80) / [1-.90\ (.80)]$$

$$= .10 / .28$$

$$= .4$$

Refer to Master Table one tailed test†(p. 36)

= .4, at 80% power requires 36 subjects

* Donner A, Eliasziw M: Sample size requirments for reliability studies. Statistics in Medicine (6):441-448, 1987

† Kraemar , Thieman: Statistical Power Analysis in Research How many Subjects? Sage Publications 1987: 32-35, 54-55

APPENDIX F
CONSENT FORM

TITLE : INTER-RATER RELIABILITY OF FORMALLY TRAINED AND SELF TRAINED RATERS USING THE EDMONTON FUNCTIONAL ASSESSMENT TOOL

INVESTIGATOR: Terry Kaasa, B.P.T.
Graduate Student, Department of Physical Therapy
University of Alberta
Phone: 436-6727 (home - evenings)
SUPERVISOR: Dr. Jean Wessel,
Professor, Department of Physical Therapy
Faculty of Rehabilitation Medicine
University of Alberta
Phone: 492 - 2812 (days)
CONSULTANT: Dr. Eduardo Bruera
Director, Palliative Care Program
Grey Nuns Community Hospital & Health Centre
Phone: 450-7044 (days)

This Consent Form, a copy of which has been given to you, is part of the process of informed consent. It should give you a basic idea of what the research project is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, you should feel free to ask. If you wish to discuss this with someone not involved in this study, you may contact Brenda Waye, Manager, Corporate Services, Caritas at 930-5892. Please take the time to read this carefully.

PURPOSE: The purpose of this study is to learn whether two physical therapists who use the Edmonton Functional Assessment Tool to assess the same patient will agree on the score.

PROCEDURE: You will be seen by two therapists at the same time. One of the physical therapists will ask you a few questions about any pain, shortness of breath, or tiredness you may have. The other therapist will watch. You will then be asked to move out of bed, and if you can, to walk or use a wheelchair. This will take about 30 minutes. You will not be expected to do anything you do not feel able to do or anything that would be unsafe for you. You may use any equipment you usually would or you may ask the therapist to help you. You may feel tired and may rest whenever you want.

CONSENT: I, _____ (please print) agree to take part in the above study which has been completely explained to me. I understand that patients on the palliative care unit are all different and that I may not be able to do all that is asked of me. I understand this study will not harm me and I will be told about my performance after both therapists have seen me.

I understand that I am a volunteer and may withdraw from this study at any time without affecting the care I am receiving. I understand my doctor is aware of and supports this study.

I understand all records have been given a code number. All information will be kept confidential. I understand that information resulting from this study may be used for educational purposes or for publication but that I will not be identified. I understand that my medical record at _____, as it relates to this study may be reviewed by Terry Kaasa. All material and data obtained from this study will be stored and may be used for future analysis without obtaining further consent from me. However, each study arising as a result of information obtained in this study will be submitted for ethics approval.

I have read and understood the information above. I sign this consent form willingly and I have received a copy of it. All questions that I had about the study have been answered. I understand that I may call either Terry Kaasa, Dr. Wessel, or Dr. Bruera at the phone numbers above if I have more questions.

_____ _____

(Name of Participant, Printed) (Signature)

_____ _____

(Name of Witness, Printed) (Signature)

_____ _____

(Name of Investigator, Printed) (Signature)

_____

(Date)

APPENDIX G

EFAT GUIDELINES ADDENDUM

Some items* of the EFAT-2 are subjective and you may find it difficult to rate them objectively. The following questions may be used to clarify these items by asking the patient to confirm your rating. You must be consistent with the questions you use.

PAIN*: What can the patient do in spite of pain?

"Generally speaking do you have pain?" No = 0. If yes then "Does it limit what you would like to do a little bit (1), considerably (2), or prevent you from doing any activity? (3)

MENTAL STATUS: Corresponds to the MMSE. If a MMSE has been done the same day you can refer to it to confirm your rating. Otherwise, you will have to ask the following:

1. What is your name? (self)
2. Can you tell me where you are? (place – hospital, unit)
3. What is the date today? (time – month, day, date, year)
4. What did you eat for breakfast? (immediate recall) OR

4a. What did you do this morning? OR

4b. What is your nurse's name? (fair only if nurse has been introduced)

5. What is the name of your doctor? (recent memory) Or

5a. Where were you before you came here?

6. When were you born? (remote memory) OR

6a. How old are you?

6b. What did you so last summer for vacation? (confirm with family)

6c. When was your grandchild born? Name? (confirm with family)

DYSPNEA: Uses a dyspnea rating scale (DRS) OR use of oxygen.
DRS: Count out loud to 15 at a rate of 2 counts/second in one breath.
Demonstrate: "I would like you to count out loud to 15 and try to do so in one breath. Like this... If this is difficult for you please take a breath whenever you need to."
Place your hand on the patient's upper chest so that you are aware of any extra breaths they take.

FATIGUE: "Generally speaking, do you feel tired?"
"Which of these statements sounds most like you?"
Refer to the descriptors and read the one(s) you feel best describe the patient. Ask the patient to comment.

MOTIVATION: What has the person been like during the assessment? This may be difficult to assess in one visit.
Have they been compliant with the assessment?
Interested?
Goals? Are they future focused? (this afternoon I am going to ...)
"Once your main concern(s) is addressed, what do you want to do?"
Ask for anecdotal evidence (must be confirmed with nurse or family):
"Are you getting up out of bed?"

- discussing your care with your doctor?"
- taking part in leisure activities?"
- enjoying visits from family and friends?"

PERFORMANCE STATUS: Global rating based on what you have just found from the previous ten items.

APPENDIX H
CONFIDENCE INTERVALS FOR INTRACLASS CORRELATIONS†

The following illustrates how to put a confidence limit on a reliability coefficient. The analysis of variance (ANOVA) table from Table 4, Chapter 4 represents results of the present study and sample 1 will be used as the example.

The reliability coefficient based on a single observation, type (1,1) can be calculated by substituting the mean square terms into the following expression:

$$R = \frac{\text{MS between-MS within}}{\text{MS between} + (k-1)\text{MS within}} \qquad (1)$$

R= the ICC, MS between = the mean squares between subjects
MS within = the mean squares within subjects, k = the number of observers

Dividing the numerator and denominator of expression 1 by MS within, R can be directly related to F.

$$R = \frac{F-1}{F+(k-1)}$$

(MS between ÷ MS within)
(83.9282 ÷ 1.2083)
New F= 69.459

The upper 95% confidence limit can be calculated by determining the corresponding F value for p = 0.025 on 30 and 30 degrees of freedom (~ 2.07) and by multiplying this by the new F value.

Upper limit = 2.07 x 69.459
= 143.78

$$R = \frac{143.78 - 1}{143.78}$$

R = .9930

The lower 95% confidence limit is determined by divding the two F values:

Lower limit = 69.459 ÷ 2.07
= 33.55

$$R = \frac{33.55 - 1}{33.55}$$

R = .9701

† Stratford, PW. Confidence Limits for Your ICC. *Phys Ther* 1989: 69 : 237

APPENDIX I
POTENTIAL SUBJECTS FOR STUDY

Sample 1

Patients interviewed	42		
Refusal to participate	4*		
Consent obtained	38		
EFAT not administered	2 →	changed mind	1
		deceased prior to rating	1

*** Patients refusing to participate**

Gender	Age	Diagnosis	MMSE	Reason for refusal
F	78	Annular Ca	?	Unable to decide
F	81	Ocular Melanoma	28/30	No research
F	78	Esophageal Ca	26/30	No research
F	74	Rectal Ca	27/30	No research

Patients not interviewed	**21**
Deceased soon after admission	7
Severe cognitive impairment	6
Language other than English	4
Deteriorating condition	3
Alzheimer's Disease	1

APPENDIX I
POTENTIAL SUBJECTS FOR STUDY

Sample 2

Patients interviewed	45		
Refusal to participate	5 *		
Consent obtained	40		
EFAT not administered	4 →	Changed mind	2
		Transfer to another facility	1
		Fatigue	1

***Patients refusing to participate**

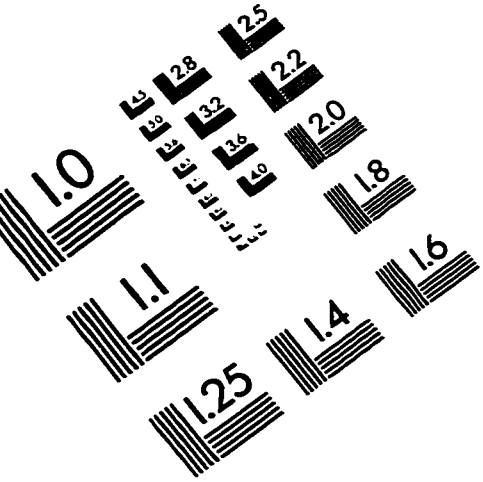
Gender	Age	Diagnosis	MMSE	Reason for Referral
M	66	Esophageal Ca	?	No research
F	86	Esophageal Ca	?	Fatigue
F	85	Pancreatic Ca	?	Fatigue
F	42	Breast Ca	?	No research
M	90	Esophageal Ca	?	Fatigue

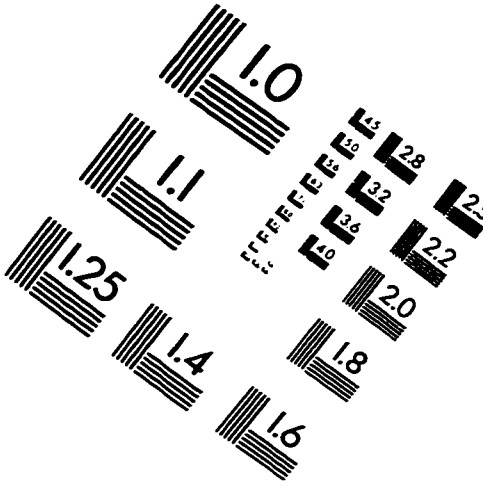
Patients not interviewed	**28**
Severe cognitive impairment	6
Deceased soon after admission	8
Language other than English	6
Deteriorating condition	3
Decreased level of consciousness	2
Transfer to another facility	1
Alzheimer's Disease	1
Discharge Home	1

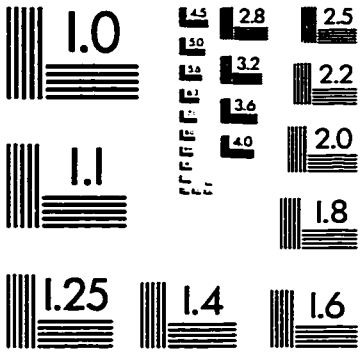
APPENDIX J
EFAT GRAPH

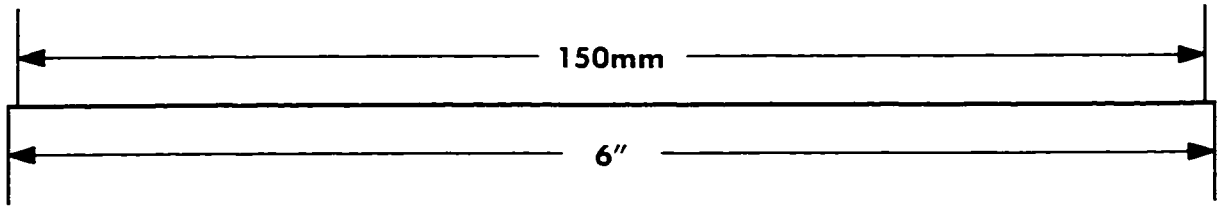
	Date							
Communication	3							
	2							
	1							
	0							
Mental Status	3							
	2							
	1							
	0							
Pain	3							
	2							
	1							
	0							
Dyspnea	3							
	2							
	1							
	0							
Balance	3							
	2							
	1							
	0							
Mobility	3							
	2							
	1							
	0							
Locomotion	3							
	2							
	1							
	0							
ADL	3							
	2							
	1							
	0							
Fatigue	3							
	2							
	1							
	0							
Motivation	3							
	2							
	1							
	0							
TOTAL								
Performance Status	3							
	2							
	1							
	0							

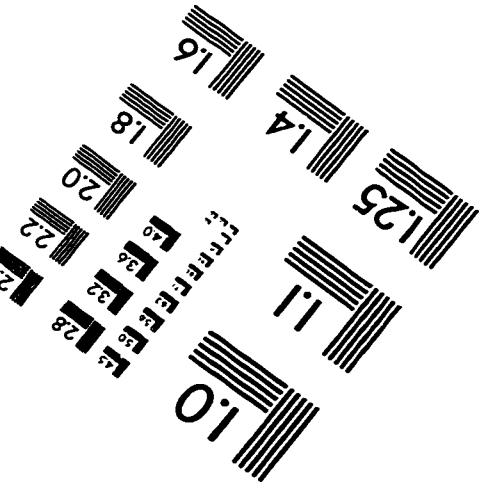
IMAGE EVALUATION
TEST TARGET (QA-3)

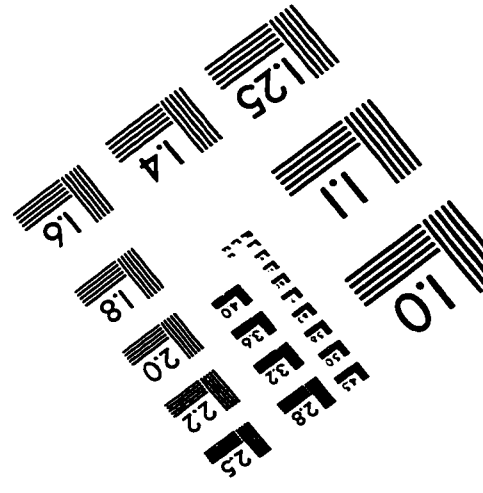












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989