A MODEL FOR ASSESSING HUMAN PERFORMANCE SYSTEMS:

An Application in Vocational Training

by

T. B. Dent

A Thesis
Submitted to the Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of
Master of Continuing Education
in the
Department of Communications, Continuing and Vocational Education
University of Saskatchewan
Saskatoon
1991

The author claims copyright. Use shall not be made of the material contained herein without proper acknowledgement, as indicated on the following page.
In presenting this thesis in partial fulfilment of the requirement for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Communications, Continuing and Vocational Education, University of Saskatchewan, Saskatoon, Saskatchewan S7N 0W0
Abstract

The purpose of this thesis was to derive from the adult education, training, and human resource development literature a model to assess human performance systems in a wide variety of situations, and to apply the model in one situation to test its efficacy. After an extensive review of the literature, a seven element performance model was developed. This model suggests the questions which must be asked and the areas to be investigated in assessing a performance system.

Woodlands Campus was selected as the site for the case study because of its unique features including an individualized competency-based learning system and computer managed testing. A multi-method approach to data collection was utilized to gather data on objectives and instructions to students and instructors, the tests and test items used, incentives to students and instructors, procedures and processes used in testing students, and the experience and qualifications of instructors. The efficacy of the model should be even more apparent when it is useful in evaluating unique performance situations.

The computerized item test bank and test generation system were systematically sampled to assess the quality of the items and the tests generated. Since performance is the goal of competency-based learning, the skills testing component of Woodlands student assessment was systematically sampled to assess quality. A sample of students was interviewed to determine how the assessment system works from the students' perspective and a sample of instructors was interviewed to assess the qualifications and procedures used by instructors in their assessment of students.
The study found that the computerized testing system, although large compared to most systems, was a simple system containing a large proportion of poor quality items most of which test only the lowest levels of knowledge. The large number of items, 73,000, provides an inadequate basis for proper testing since it only results in a ratio of 2:1 for total items from which to draw, whereas, experts suggest a 10:1 ratio is necessary. The performance testing ranged from sophisticated to simple with the largest proportion of tests being simple, low-fidelity tests. Recommendations for improvements were made on the basis that three stages were necessary commencing with improvement of the organizational climate, followed by the improvement of staff knowledge and skills which are necessary to improve the student evaluation system. Although improvement of the student evaluation system is the ultimate goal, it is not likely to be achieved by instructors lacking the necessary educational knowledge, skills, and motivation to acquire them.
Acknowledgements

The writer wishes to express his gratitude to all those who contributed to this thesis. I am especially grateful to Dr. A. Blunt, my thesis supervisor, for his patience, understanding, guidance, and encouragement throughout the two years it took to complete this thesis. I am grateful for the support and advice of my committee members, Dr. A. Yackulic, and Dr. M. Collins. I would like to thank Dr. F. B. Brown who acted as my initial advisor and Professor W. McKay and his voc-ed students for their assistance in rating samples of data. I would also like to thank the students, instructors, and administrators of S.I.A.S.T. Woodlands Campus who cooperated with me in the collection of data and interviews. Finally, I would like to thank my family, Laurie, Bruce, and Nancy, for their perseverance and forebearance throughout the entire process.
# Contents

Abstract
Acknowledgements
List of Tables
List of Figures

## Chapter 1. Introduction

- Statement of Goal and Objective
- Statement of the Problem
- Delimitations
- Assumptions
- A Case Study as an Example of the Application of the Model
  - Historical Background of Woodlands Campus
  - Description of Woodlands' Student Evaluation System
  - Definitions and Terms
  - Plan of the Study

## Chapter 2. Literature Review

- The Problem
- The Search For and Evolution of a Model
  - The Four Element Model
  - Gilbert's Model
  - The Seven Element Performance Model
- The Application and Relevance of the Model
- Performance Evaluation
  - Definition
  - Value
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>30</td>
</tr>
<tr>
<td>Methods</td>
<td>32</td>
</tr>
<tr>
<td>Assessment Centres</td>
<td>37</td>
</tr>
<tr>
<td>Rationale</td>
<td>37</td>
</tr>
<tr>
<td>Description</td>
<td>38</td>
</tr>
<tr>
<td>Validity</td>
<td>41</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>42</td>
</tr>
<tr>
<td>Educational Measurement</td>
<td>43</td>
</tr>
<tr>
<td>Norm-referenced and Criterion-referenced</td>
<td>44</td>
</tr>
<tr>
<td>Testing</td>
<td>45</td>
</tr>
<tr>
<td>Essential Characteristics</td>
<td>46</td>
</tr>
<tr>
<td>Types of Testing</td>
<td>47</td>
</tr>
<tr>
<td>Performance Testing</td>
<td>48</td>
</tr>
<tr>
<td>Written Tests</td>
<td>49</td>
</tr>
<tr>
<td>Computer-assisted Testing</td>
<td>50</td>
</tr>
<tr>
<td>Performance Assessment as a Performance</td>
<td>52</td>
</tr>
<tr>
<td>Knowledge</td>
<td>52</td>
</tr>
<tr>
<td>Skills</td>
<td>53</td>
</tr>
<tr>
<td>Motives</td>
<td>54</td>
</tr>
<tr>
<td>Data</td>
<td>54</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>55</td>
</tr>
<tr>
<td>Incentives</td>
<td>55</td>
</tr>
<tr>
<td>Climate</td>
<td>56</td>
</tr>
<tr>
<td>Application</td>
<td>57</td>
</tr>
<tr>
<td>Competency-Based Learning</td>
<td>57</td>
</tr>
<tr>
<td>Definition of Competence</td>
<td>57</td>
</tr>
<tr>
<td>Definitions of Competency-Based Learning</td>
<td>60</td>
</tr>
<tr>
<td>What is Competency-Based Learning?</td>
<td>61</td>
</tr>
<tr>
<td>Problems and Criticisms of Competency</td>
<td>64</td>
</tr>
<tr>
<td>Measurement of Competence</td>
<td>69</td>
</tr>
</tbody>
</table>
Chapter 3. Design and Methodology

Process and Product Evaluation
Process Evaluation Using the Performance Model
A Multi-Method Approach
Assessing the Computer Item Bank
  Rationale.
  Method.
Assessing the Computer Generated Tests
Assessing the Skills Tests
Assessing the Instructional Objectives
Interviews

Chapter 4. Analysis of Data

Computer Test Bank Item Analysis
Analyzing the Computer Tests
Analyzing the Skills Tests
Analyzing the Objectives
Analyzing the Student Interviews
Analyzing the Instructor Interviews
Summary of Data Analysis

Chapter 5. Conclusions and Recommendations

Conclusions
Knowledge Testing
  Instrumentation – the computer testing system.
  Data – instructions and feedback.
  Incentives.
Performance Testing
  Instrumentation.
Data - instructions and feedback. 112
Incentives. 112
Motivation and Attitude Measurement 113
Instrumentation, data, and incentives. 113
Recommendations 113
Improve the Organizational Climate 113
Improve the Instrumentation for Instructors 116
Improve the Incentives for Instructors 117
Improve the Data to Instructors 117
Improve the Human Element 118
Improving the Student Assessment System 120
Improve the computer testing system. 120
Improve the performance testing system. 121
Improve the assessment of student attitudes. 122

Chapter 6. Summary 123
The Study Goal 123
The Study Objective 123
Assessing Human Performance 124
The Environment 124
Data. 124
Instrumentation. 124
Incentives. 125
Climate. 125
The Personal Factor 125
Knowledge. 125
Skills. 126
Motives. 126
Application of the Model 126
Methodology 127
Analysis of Data 128
The Computer Testing System 128
Performance Testing 129
Objectives 129
Student Interviews 130
Instructor Interviews 130
Conclusions and Recommendations 130

References 133

Appendix

A  Letter to Dr. Brown 149
B  Letter to Mr. Peter Kalyn 151
C  Letter to Mr. Peter Kalyn 152
D  Letter from Mr. Peter Kalyn 154
E  Student Interview Schedule 155
F  Instructor Interview Schedule 159
List of Tables

Table 1 Characteristics of Interview Sample and Campus Student Population 85
Table 2 Multiple-Choice Item Flaws 87
Table 3 Computer Test Analysis Summary 90
Table 4 Summary of Skills Test Analysis 93
Table 5 Summary of Analysis of Instructional Objectives 94
Table 6 Instructors Educational and Experience Background 99
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Four Element Performance Model</td>
<td>18</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Gilbert's Performance Model</td>
<td>19</td>
</tr>
<tr>
<td>Figure 3</td>
<td>The Human Factor</td>
<td>21</td>
</tr>
<tr>
<td>Figure 4</td>
<td>The Environmental Factor</td>
<td>23</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Seven Element Performance Model</td>
<td>26</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

Statement of Goal and Objective

The goals of this study were to develop a model of excellence in performance evaluation and demonstrate its application in one practical situation. The model was developed after an extensive review of available literature in educational measurement, performance evaluation in business and industry, and the certification of professionals. Additionally literature was reviewed in the areas of competency-based education and computer-assisted testing because of the relevance of these topics to the study.

The objective of the study was to compare the Saskatchewan Institute of Applied Science and Technology (S.I.A.S.T.), Woodland Campus student evaluation system with criteria suggested by the developed model and derived from the literature indicative of the present state of the art in performance evaluation. The student evaluation system was assessed by utilizing the performance model to suggest questions about each of the elements of the performance system in question.

Statement of the Problem

A theoretical model of human performance assessment constructed systematically from a review of the literature will provide a basis for revealing inadequacies in the existing evaluation system at Woodland and will identify areas in need of improvement.
Delimitations

This study is concerned with the methods and techniques utilized to measure, judge, and certify human vocational performance as competent or not yet competent. It is concerned with the mechanics of the system and will not consider all of the philosophical questions as to the efficacy of competency-based learning or whether or not performance evaluation should be done.

One of the basic problems of any assessment system is the problem of false positives and false negatives. That is, those certified as competent who are not really competent (false positives) and those who are not certified as competent when they actually are (false negatives). A good evaluation model should lead to a system which minimizes this problem and, although this is one reason for doing this study, I will not attempt to assess whether or to what degree Woodland's evaluation system actually accomplishes this end.

The literature review will mention biases such as racial prejudice, and sexism which can affect performance evaluation; no attempt will be made in this study to ascertain whether or not these factors are operational within the context of Woodland's student evaluation system. Similarly, no attempt will be made to assess the appropriateness of the level of basic skills (reading, math, etc.) which may be necessary to successfully complete Woodland's programs since this is more a function of program development than evaluation. Each of these matters could be the subject of a separate study.

The individualized and relatively unstructured nature of student attendance and evaluation at Woodland combined with the limited time and resources of one evaluator imposed restrictions on the nature of the data collection for this study. For example, a
preferred method of gathering data for the analysis of skills testing was to observe a sample of skills tests being administered by instructors. Because most skills tests are administered "on request" by the student and not scheduled well in advance, selecting a sample for observation was not possible. Therefore, other methods were necessary. Questions were included in the interview schedules for both student and instructor interviews to determine how the skills tests were conducted.

Assumptions

People make frequent judgements about human performance and competence. They always have, and they probably always will. For example they judge whether the person who repaired their car did the job well or whether the teachers who teach their children are competent or not. We make such judgements often and usually on the basis of poor and insufficient data. Therefore, the issue is not whether we can measure performance or certify competence; the real issue is how well can we do it! For the purposes of this study it is assumed that human performance can be measured; it is acknowledged that there are factors which can operate to reduce the accuracy of such an assessment but the development of a good evaluation system helps to reduce and control these factors.

A Case Study as an Example of the Application of the Model

One method of demonstrating the value of the developed model is to apply it to a particular situation. Although the model could be applied in many situations, there are a number of reasons for selecting the relatively new Woodland Campus of S.I.A.S.T. as the case study. The competency-based vocational training programs developed at Woodland Campus (originally the Northern Institute of
Technology) are unique in their dependence on a computer-managed learning and testing system. (Rusnell and Collins, 1989) Although competency-based educational systems have been implemented in many places and in many differing educational situations, and although computer-managed learning systems and testing systems have also been widely implemented, there does not appear to be any existing model on which to evaluate a learning environment which incorporates all of these factors.

An evaluation of the Woodland's CBE system by Rusnell and Collins (1989) was completed two years after the campus opened. This evaluation study encompassed all aspects of this CBE system and raised questions about the student evaluation system but it did not examine the system in detail. The evaluators suggested that, "Woodlands is nearing a point where it can begin to refine its testing procedures, ..." and then suggested some areas for investigation and improvement including the need to improve the computer test item bank and to develop other testing procedures.

Thus, close at hand was an education facility with unique features which depended on the ability to assess human job oriented performance and in need of study to assess its ability to assess this vocational performance. The fact that the developed model provided a framework to evaluate the unique features of Woodland's computer-managed competency-based system would serve to illustrate its utility more than if it were used to assess a more traditional delivery system, such as that utilized at other vocational training institutes. Further, the fact that the assessment could be valuable to the institution in improving its system made the study all the more valuable and interesting.
Historical Background of Woodland Campus

The Saskatchewan Institute of Applied Science and Technology (S.I.A.S.T.) is a provincial post-secondary education and training institution which was created by legislation on January 1, 1988 by amalgamating the province's four technical institutes and six community colleges. Woodland Campus (S.I.A.S.T.) in Prince Albert, Saskatchewan, is one of four campuses which provides a variety of post-secondary training and educational opportunities and was formed by amalgamating the Northern Institute of Technology, the Prince Albert Regional Community College, and the Meadow Lake Vocational Training Centre. Woodland Campus serves the northern half of the province of Saskatchewan through its main campus and four additional locations in Prince Albert, the Meadow Lake vocational centre, and fourteen other northern locations which are linked by computer to the main campus. (Rusnell and Collins, 1989)

The former Northern Institute of Technology was planned and constructed in the early 1980's and commenced operation in September, 1986. It was intended to be a unique training institution with programs designed and delivered through an individualized competency-based format utilizing a computer managed instructional monitoring and evaluation system. The specific features of this institution were to be: individualized modular instruction; year round operation with continuous intake and exit; extended daily hours of operation; possibility of part-time enrollment; few prerequisite qualifications for entry; the possibility of challenging knowledge and performance tests at any time; and the delivery of programs at a distance. (Woodward, 1986)
Description of Woodland's Student Evaluation System

Proponents of CBE claim that what differentiates this approach from "traditional education" is the emphasis on ability to actually perform or apply learning as opposed to just knowing about that performance. (Blank, 1982; Klemp, 1979) This emphasis on performance instead of only on knowledge alone implies the ability to measure performance, which is much different from measuring knowledge about performance; thus, traditional testing and measurement schemes are not appropriate for a competency-based educational program. According to Blank (1982, p5) one of the major criteria which differentiates competency-based programs from traditional programs is that CB programs "Require each individual trainee to perform each task to a high level of proficiency in a joblike setting before receiving credit for attaining each task."

Add to this the fact that the instructional programs at Woodland are individualized with students being able to proceed at their own pace and the requirement for an evaluation system that is also individualized is evident. Since students must successfully complete each module before proceeding to the next, then they must be able to take the required test when they are ready if they are to proceed at their own pace.

There are approximately 3500 different competencies within the 40 programs delivered by Woodland, all of which have to be tested for prerequisite knowledge and actual performance. Some examples of competencies at Woodland are: (1) Apply modified bitumen membrane, from the roofing program; (2) Make pie filling, from the commercial cooking program; (3) Lighten and tone hair, from the cosmetology program; and (4) Describe print media as a socializing agent, from the early childhood development program. On any given day there will be about 500 student requests for tests, and
since some of them are retests, it is necessary to be able to
generate, administer and score a great many different tests
quickly, on demand across all program areas. This is the reason
for utilizing a computer to generate and score the knowledge tests
and manage the record keeping necessary to keep track of the
progress of 1100 students at various points in the programs.

There are two types of tests utilized at Woodland, (1)
performance tests (called skills tests) and (2) prerequisite
knowledge tests which are basically "paper and pencil" types of
tests which are delivered via a computer terminal. Most of these
latter types of tests are composed of multiple choice items of which
there are about 73,000 stored in the computer test bank. The
computer generates a test by selecting a predetermined number of
items at random from a specified set of items organized by
competency in the test bank. When students feel ready to
challenge the knowledge test for the competency being attempted,
they go to the testing centre and make a request to take the
relevant test. If their fees are paid and they have completed all of
the required prerequisites students are assigned to a computer
terminal through which the test is delivered and they complete the
test. The computer scores the test and the student can receive
the score from another terminal outside the testing centre within a
few minutes of its completion. If the student wishes to review the
test an instructor is requested to call up the actual test and
responses on a terminal in the instructors office.

The knowledge test completed on the computer is a
prerequisite to being allowed to take the skills test for the
particular competency being learned. If the student fails the
computer test (a score of less than 80 percent), he or she must
return to the original instructional materials for further study or
seek assistance from an instructor. When ready, the student can
try a knowledge test again. The computer will generate another
test in the same manner as it generated the first. A student who fails three consecutive tries is "locked out" of the system by the computer until an instructor or teaching assistant unlocks it for them. This is allowed after the student has given a good explanation for the previous unsuccessful attempts.

Once the knowledge test has been passed, the student may request the skills test, which is then administered and scored by an instructor. The results of this test are entered into the computer by the instructor. Parallel versions of the skills tests are kept for subsequent challenges if the student's first challenge is not successful. The computer retains the results of all tests taken but only generates and scores the knowledge tests.

**Definitions and Terms**

The following terms appear throughout this thesis as they do in the literature. Since terms are not always defined in the same way by different authors, the definitions I have adopted are provided here.

(1) *Performance assessment* is the process of systematically gathering data on an individual's performance and comparing the data to some established standards of adequacy or excellence for making decisions about the individual. (Berk, 1986; Guion, 1986; Sokol & Oresick, 1986)

(2) "*A job analysis is a set of observational procedures which yield a description of what a job incumbent does, as well as how and why the incumbent does it."* (Aldag & Brief, 1979)
(3) "A task is a unit of work performed by one person as part of his/her total job responsibilities." (Hindes, 1976 p.13) A task has a definite beginning and ending and results in a useful product or service. (Levine, 1983)

(4) A function (or duty) is a collection of tasks related to the accomplishment of a major responsibility of the job.

(5) Competence is a judgment made about a measure of human performance while performance is a matter of how well a certain function is carried out. The assessment may be on the product, the process or both. (Jones and Whittaker, 1975; Shepard, 1979)

(6) "Competence-based education tends to be a form of education that derives a curriculum from an analysis of a prospective or actual role in modern society and that attempts to certify student progress on the basis of demonstrated performance in some or all aspects of that role. Theoretically, such demonstrations of competence are independent of time served in formal educational settings." (Grant, 1979 p.6)

(7) Competency-based vocational training is a systematic training scheme based on programs developed from a thorough job analysis; delivered in a variety of ways congruent with principles of mastery learning; and utilizing high fidelity measurement techniques to determine the competence of students before certifying the same.
(8) "Mastery learning is an instructional methodology which utilizes criterion-referenced testing in a formative manner to ensure that all students master the parts of the curriculum to a high degree before proceeding to learn the next part." (Guskey, 1985)

(9) High-fidelity measurement is measurement in which the reproduction of cues and responses is identical to or a close approximation of an on-the-job situation. (Jones and Whittaker, 1975)

(10) Criterion-referenced testing is a testing technique which attempts to measure student achievement against predetermined standards of achievement. (Grant and Kohli, 1979; Popham, 1978)

(11) Norm-referenced testing is a testing technique which attempts to measure student achievement in relation to other students achievement on the same curriculum. The purpose of this testing is to differentiate between students. (Nitko 1983, Gay 1980)

(12) Computer-assisted testing refers to the use of a computer to do any or all of the following functions: generate tests, administer tests, and score tests.

(13) Multiple-choice tests are constructed from items having a stem which asks a question or poses a problem followed by a number of possible answers, only one of which should be correct. The remaining answers are called distractors and should present a possible answer which is plausible to the uninformed (Gronlund, 1981). These tests are often called objective tests because they
are scored by using a supplied answer key thus removing subjective judgement from the scoring process.

(14) *Equivalent forms* also called alternate forms or parallel forms are different forms of a test "built to measure the same abilities (that is they are built to the same set of specifications) but constructed independently ..." (Gronlund, 1982 p.133)

(15) *Test specifications*, sometimes called test blueprints, are a testing plan which specifies the content domain, level of learning, and sample performances as evidence of satisfactory achievement. (Gronlund, 1982)

(16) *False positive* is a category of persons who achieve a passing mark on a test or were certified as competent when, in fact, they are not knowledgeable or competent.

(17) *False negative* is a category of persons who failed a test or were not certified as competent when, in fact, they are competent or knowledgeable.

(18) *Validity* refers to the question of whether or not the results of a test can be interpreted in the manner we wish to use them. "Are we measuring what we think we are measuring?" (Kerlinger, 1986) Kerlinger (1986) discusses three types of validity: (a) content validity, (b) criterion-related validity, and (c) construct validity.
(19) "Content validity is the representativeness or sampling adequacy of the content -- the substance, the matter, the topic -- of a measuring instrument." (Kerlinger, 1986 p.417)

(20) Criterion-related validity is sometimes discussed under two separate headings: (a) concurrent validity, and (b) predictive validity.

(i) Predictive validity is the degree to which a test enables one to predict future success on some other criterion, such as successful performance on a job. According to Kerlinger (1986), "A test high in criterion-related validity is one that helps investigators make successful decisions in assigning people to treatments..." (p.419)

(ii) Concurrent validity is a measure of the degree of success current successful performers would have on a test. In other words, if successfully practicing mechanics did not do well on a test designed to measure competence as a mechanic, then the test would have low concurrent validity. Concurrent validity is not a substitute for predictive validity; however, if a test has poor concurrent validity, then its predictive validity would necessarily be doubtful since it would cast doubt on the criteria selected to define successful practice.

(21) Construct validity refers to the degree to which some psychological construct can be used to explain individual variances in test scores. In short, what does the test really mean? (Kerlinger, 1986)
Some writers also mention "face validity." Face validity means that the test is generally seen on the basis of its content items as being valid for its purpose. Although, expert psychometricians point out that face validity is an ersatz validity, Stewart and Stewart (1981) claim that the public relations value of face validity should not be underestimated.

"Reliability is the accuracy or precision of a measuring instrument." (Kerlinger, 1986 p.405) An accurate instrument must produce stable results; it must produce the same or similar score each time it is used on the same object or event. A coefficient ranging from 1.0 to 0.0 is often used to express the reliability of an instrument.

Intra-rater reliability refers to the ability of a rater to make the same judgement about an observed performance or category of items in a test each time it is rated. If a teacher scored a selection of essays three times at intervals of several days or weeks, without access to or recollection of the previous scores, and produced very similar scores on each occasion, then the intra-rater reliability of that teacher's judgements would be high for this task.

Inter-rater reliability refers to the degree to which the scores or judgements of different raters would agree when rating the same performance or categorization of items. A coefficient ranging from 1.0 to 0.0 may be used to express the degree of agreement or a percentage of agreement may be calculated.
Test objectivity is achieved when the results of a test are obtained through means which are independent of judgements by the examiner or test scorer. Different appraisers would all arrive at the same results given the same testing situation because they would all utilize the same answers or criteria and definitions which would have been defined before the assessment.

Sensitivity refers to the ability of a test to differentiate accurately between different levels of performance. A test with high sensitivity can differentiate between small observable differences in performance while a test with low sensitivity could only differentiate between large differences in performance.

Functionality is concerned with the practicality of the measurement scenario. All testing situations are limited by available resources and therefore what it is practical to do, however, a practical test which is not valid or reliable has no value.

Individualized instruction is instruction which is designed for the individual student rather than for groups. It usually involves flexible pacing and may offer flexible content presentation and sequencing. (Miller and Rose, 1975)

Knowledge as defined by Bloom et al (1956):

... "involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material."
(31) *Skills* may be psychomotor, cognitive, affective, or a combination of these three. Psychomotor skills refers to motor skills, from fine to gross manipulation of physical objects, and necessary to manipulate tools, operate machinery, et cetera. Cognitive skills include the top five levels of cognitive functioning specified in Bloom's taxonomy: comprehension, application, analysis, synthesis, and evaluation. Affective skills refers to human abilities to empathize, sympathize, and feel with or about others and are necessary to function competently in occupations where direct service to the public are involved. These skills have been defined by Krathwohl et al (1964).

(32) *Attitudes* will be defined according to the taxonomy developed by Krathwohl et al (1964). Attitudes encompasses levels 2.2 through to 4.1 of this taxonomic category and includes affective responses from willingness to respond through preference and commitment to conceptualization of a value.

(33) Performance evaluation is "the measurement of the effectiveness of job-related behavior." (Guion, 1986)

**Plan of the Study**

Chapter 2 develops the performance model used for this study. It reviews the literature on performance evaluation and measurement in business and education and relates this literature to the model. A review of the literature defining competency-based learning and the arguments for and against CBL by its proponents and critics is included because it is relevant to the particular institution, Woodland Campus, which was used for the case study. Chapter 3 describes the multi-method approach used to gather the
data for the analysis of the student evaluation system at S.I.A.S.T.'s Woodland Campus and presents a rationale for the methodology. Chapter 4 analyzes the data collected and presents the results obtained from each method: (1) the sample of test items, (2) the sample of computer generated knowledge tests, (3) the sample of skills tests, (4) the sample of objectives, (5) the interviews with a sample of students, and (6) the interviews with a sample of instructors. Chapter 5 presents the conclusions arrived at by incorporating the results of the data analysis into the performance model and the recommendations for improvement of the student evaluation system at Woodland. Chapter 6 summarizes the study.
Chapter 2. Literature Review

The Problem

Although much has been written about the importance of assessing human performance, the benefits to be obtained from such assessment, and from individual components of such assessment (instruments, interviews, biases, etc.), little has been done to integrate these various aspects into a comprehensive model which would guide the process of analyzing and assessing human performance. Thus most performance assessment schemes are fragmented with an emphasis, if there is one, on whatever limited aspect the assessors happen to have knowledge of, or particular interest in. The lack of a systematic approach results in performance appraisals which are of limited usefulness since they often lack the basic characteristics of a good measurement system such as validity and reliability, and may even be unfairly biased against the persons being assessed. A systematic approach by itself, is not a guarantee of a quality assessment. Such an approach must be guided by a comprehensive understanding of the phenomenon being measured, its various components, and their interactions. For this reason a comprehensive model which provides a reasonable explanation of the elements and interactions which result in human performance is needed.

The Search For and Evolution of a Model

The next section will describe the evolution of the seven element performance model which became the model of human performance on which this study is based. This model evolved as a direct consequence of the fairly extensive literature review carried out as a central part of this thesis.
The Four Element Model

An extensive search of the literature in the disciplines of education and commerce and to a lesser extent in medicine and law resulted in a great many books and articles being identified which dealt with the evaluation of students, the certification of professionals, and the assessment of employees. However, no overarching framework to link the various aspects of these activities together was identified. The concept of a three element model of human performance was developed from the education literature which suggested that: (1) knowledge, (2) skills, and (3) motivation, were the three interacting elements which make up human performance. Although educators in the school system seemed to be primarily concerned with these three areas, adult educators such as Malcolm Knowles (1980,1984), Cross (1986), and Wlodkowski (1988), also emphasize the effects of the environment on the performance of adult students. Thus a four element model which consisted of two basic factors, (1) the human factor, and (2) the environment seemed to be advisable. The human factor could be divided into the three elements previously identified while the environment factor had only one element.

Figure 1

Four Element Performance Model

<table>
<thead>
<tr>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
</tr>
<tr>
<td>Skills</td>
</tr>
<tr>
<td>Motivation</td>
</tr>
</tbody>
</table>
One advantage of the four element model is that it permitted the integration of adult education human resource development literature and literature from the discipline of commerce which dealt in more detail with the effects of the environment on human performance than did most of the other disciplines.

**Gilbert's Model**

The literature search identified the work of Thomas Gilbert (1978, 1982a, 1982b) which proposed a more elaborate six element model and a detailed plan for the practical application of the model in the assessment and improvement of human performance. Gilbert's model consisted of two basic factors, (1) the human factor, and (2) the environment, with each being subdivided into three corresponding elements. The three elements of the human factor are: (1) knowledge, (2) skills or abilities, and (3) motives. The three corresponding elements of the environmental factor are: (1) data, (2) instrumentation, and (3) incentives.

**Figure 2**

**Gilbert's Performance Model**

<table>
<thead>
<tr>
<th>Environmental Factor</th>
<th>Information</th>
<th>Instrumentation</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Instruments</td>
<td>Incentives</td>
<td></td>
</tr>
<tr>
<td>Human Factor</td>
<td>Knowledge</td>
<td>Capacity</td>
<td>Motives</td>
</tr>
</tbody>
</table>
The data which are provided by the environment serve two purposes: (1) to provide direction as to what is to be accomplished, and (2) to provide feedback as to how well it is being accomplished. Instrumentation includes the physical plant, machinery and tools, as well as processes and procedures necessary for the accomplishment required. Incentives include salary, bonuses, praise, recognition, etc. which are used to reward and thus motivate the person to produce a "worthy accomplishment" (Gilbert, 1978). Even though Gilbert's model (as depicted) is represented with solid lines between the elements, he makes it clear in his text that the elements of each factor overlap to some extent and that it is the interaction between the human factor and the environment which produces performance.

**The Seven Element Performance Model**

Gilbert's model is not complex, yet it explains a great deal about human performance. It is practical and provides a functional method for assessing human performance as well as indicating reasons for doing so. However, there is a growing volume of literature which does not easily fit this model. This literature which has grown rapidly in the last ten years and contains the writings of Drucker (1974), Ouiche (1982), Peters and Waterman (1984), Peters and Austin (1985), and Peters (1987), among others, deals primarily with the effects of the psychological environment or what is commonly called the organizational climate on the productivity of workers in organizations. The interaction between the personal factor and this seventh element appears to be indirect; that is, the organizational climate is perceived by people through their interactions with the communications, facilities and processes imposed by the organization, and the incentives offered by it.
As a result of this insight, an enhanced model of human performance was derived by expanding Gilbert's model to include a new element. The resulting seven element model is a little more complex, but is more comprehensive, thus providing a more satisfactory explanation of the complex phenomenon of human performance. This model retains the two major factors of the Gilbert model: (1) the human factor, and (2) the environmental factor.

Figure 3

**The Human Factor**

![Diagram of the Human Factor]

The human factor is composed of three elements, (1) knowledge, (2) skills, and (3) motives, which are represented in the figure by the three overlapping circles. Overlapping circles are used as in Venn diagrams to indicate that the three elements cannot be totally separated into discreet entities. The three elements are defined as follows: knowledge as cognitive functioning at the lowest level of Bloom's taxonomy, also called knowledge; skills as all the levels of the psychomotor taxonomic domain, the first four levels of the affective taxonomic domain, and levels two to six of the cognitive taxonomy. An example of overlap between these first two elements
is the second level of the cognitive taxonomy commonly called "comprehension" or "understanding" which could reasonably be included in the knowledge element in many cases. Motives, the third element, is defined as the individuals motives or reasons to exert effort in the performance of a task. Since motives are effected by one's attitudes and beliefs, the overlap between this element and the affective domain included in the skills element is evident.

The three elements of the human factor interact to produce human behaviour. Knowledge and skills interact to produce abilities, what people are capable of doing but not necessarily what they choose to do. The following is a more symbolic representation of this relationship.

Knowledge * Skills --> Abilities

The asterisk, which is commonly used in computer languages to indicate a multiplicative function, is used here to convey the impression that this is not a simple additive function. In other words the result may be more than just the sum of the parts and in the same manner an arrow is used instead of an equals sign. Thus, the proper reading of this equation would be: knowledge interacting with skills yields abilities.

Motives, the third element, determines the amount of effort which will be expended on a task and this effort interacts with abilities to produce human behaviour.

Abilities * Effort --> Behaviour

However, behaviour is not performance! This common misconception is at the root of a great many faulty performance evaluation systems. (Gilbert, 1978) Behaviour does not become performance
until it interacts productively with an appropriate environment.
Gilbert's (1978) metaphor of a fish is an illustration of this principle.
We see a fish swimming in the water; swishing its tail and fins and
slipping seemingly effortlessly through the water. If we can grab
the fish and throw it up on the shore, it will still move its tail and
fins, perhaps even more energetically than before, but, it will not
be progressing very far. The behaviour is still there but now the
environment is no longer appropriate and performance is lost.

Figure 4

The Environmental Factor

The second factor, the environment, is subdivided into four
elements which are represented in the drawing as three
overlapping circles all of which are overlapped by the
organizational climate represented by an oval (See figure 4).
Again the overlap indicates that clear boundaries between the
elements do not exist. Data, the first circle can be defined as the information provided to the worker to show (a) the direction in which behaviour should be oriented to accomplish the desired end (clear and explicit objectives) and (b) feedback to let the workers know how well they are actually performing. The second circle "facilities" represents all the physical plant, machinery, tools, materials, etc. required for the performance of the task. Gilbert (1978) also included procedures and processes in this element which he called instruments, however, it can be seen that this is an area in which this element overlaps with the previous element, data. The third circle, "incentives," represents both monetary and non-monetary (hard and soft) inducements to motivate workers to behave in the direction and degree desired by the organization. Since facilities such as a better office can be used as an incentive, the overlap between these two elements is also evident.

The large oval at the top of the drawing which intersects the other three circles is labelled the "organizational climate". Steers and Lee (1983) define organizational climate as the personality of an organization as seen by employees. This element is often considered an intangible but contemporary management theorists and practitioners have been vociferous in pointing out its overall importance and the many small details which contribute to the development and maintenance of organizational climate. Some of these are the organizational structure, authoritarianism, trust, and the basic philosophy of management towards employees and business in general. The employees perception of this climate is primarily formed through interactions with the data, facilities, and incentives elements since all of these elements are significantly affected by the organizational climate. As an example, if the organization is rigidly authoritarian (MacGregor's type "X"), the lines of communication will be rigidly maintained, the facilities will be supplied and maintained without consultation with those utilizing
them, and the use of non-monetary (soft) incentives is likely to be non-existent. In general, workers will be perceived as unreliable, immature, and even dishonest.

Depending on the particular circumstances, there are a great many interactions possible between the seven elements of the model. Some of these will be quite subtle and difficult to discern or influence while others such as the interactions between the three elements of the personal factor which we have already discussed, are much more evident and have been studied more thoroughly. These interactions are important to the application of the model as a practical strategy for devising an evaluation. The environmental element "data" interacts primarily with the personal element "knowledge" and this interaction can be characterized as communications (See figure 5). In a similar manner the interaction between the environmental element facilities and the human factor element skills can be considered as the science of ergonomics. The third major interaction occurs between the environmental element "incentives" and the human factor element "motives." The expectancy-valence model of motivation can be used to characterize this interaction labelled "motivation”.

Although Gilbert (1978, 1982a, 1982b) is very clear on the overall interaction between behavior and environment,

( Behaviour * Environment --> Performance )

he did not believe it necessary to break this interaction down into separate and somewhat parallel interactions as the expanded seven element model does. The reasons for adopting the more elaborate model are to improve its explanatory capability by linking it to
Figure 5

The Seven Element Performance Model

Organizational Climate

Environmental Factor

Data
Facilities
Incentives

Communications
Ergonomics
Motivation

Human Factor

Knowledge
Skills
Motives
existing knowledge and to provide guidance to performance assessors. One might find that the knowledge and the data elements in a performance situation are both positive and yet there still is a communications problem. The answer may then lie in an in-depth study of communications theory and research. It can be seen that the model can be applied as a single stage model in the manner of Gilbert, or it can be applied as a two stage model where in depth analysis is necessary. In the first stage the assessor examines each of the elements of the model to recommend correctives to those elements which appear deficient. If the elements appear to be trouble free before or after they have been corrected and there is still a performance deficiency, then the second stage or in-depth analysis can be conducted.

The Application and Relevance of the Model

Gilbert (1978, 1982a, 1982b) applies his model to the task of assessing performance by deriving what he calls a "question model" from the performance model. That is, the performance model is used to suggest relevant questions which are asked of the situation under assessment in order to determine the status of the various elements of the performance model. The seven element model adopted in this study can be utilized in the same manner for a single stage analysis. If the results of the first stage indicate the requirement for an in depth analysis, then the model is used to suggest questions relevant to the major interactions of the model. This is a considerable improvement on most assessment schemes which are limited to consideration of only the human factor elements. Since behaviour is only one of two factors which yield performance, assessing behaviour alone may not yield an adequate measure of performance. Few people would consider assessing
performance solely on the basis of an environmental assessment and yet, many assess performance strictly on the basis of weak measures of human behaviour. On the other hand, many writers point out the inadequacy of assessing only outcomes and Gilbert (1978) agrees that such measures are not adequate for all purposes. This is particularly true in training situations involving persons who are not knowledgeable and skilled. In these situations it is important to assess behaviour in order to assist the individual to learn proven and productive skills. However, once the person is competent, it is probably advisable to place emphasis on the assessment of accomplishments rather than on the methods of achievement.

This performance model is very generic and can be applied in a wide variety of situations including education, business, industry, military, government, etc.. The model is applicable to both individuals and groups and at various levels of a wide variety of organizations. The specific questions asked and the methods of data collection will vary with the situation but the elements of the model and their interrelationships will not change. Therefore, the model is useful and practical for the assessment of existing performance situations and the development of new performance systems.

Performance Evaluation

The literature on human performance assessment is broad and voluminous. Every specialty area whether business, education or professional has its own performance literature seemingly on the assumption that problems of human performance are unique in each area. Few writers cross the boundaries of their own discipline to seek enlightenment in other fields. For this study a broad search has been undertaken through materials from commerce, education,
and other disciplines. In the process I have discovered that the problem is essentially the same in all areas. How do we create a comprehensive human performance assessment system which is valid, reliable, objective, sensitive, and functional?

Definition

The terms performance assessment, performance evaluation, and performance appraisal are used interchangeably in the literature. Berk (1986) says performance assessment is a "systematic process" but most evaluations are done on a periodic rather than an ongoing basis and are far from systematic. Guion (1986) states that it is a "measure of effectiveness," yet most evaluations utilize indirect measures such as behaviour or attributes and thus are useless for measuring effectiveness. Sokol and Oresick (1986) require "a comparison of performance to standards", however, few performance standards are stated in such objective terms that they could be recognized and agreed upon by different raters let alone ratees. It would seem, therefore, that current practice falls far short of the objective ideals in the definitions. Granted, a good performance evaluation scheme must be functional, but to excuse an unsystematic, subjective, and unreliable evaluation on the basis that it is "practical" is to be blind to the value of doing assessments in the first place.
The value of performance appraisals does not come from doing them, it comes from the quality of the decisions which can be made as a result of reliable, valid, and objective data which such an exercise should yield. According to Landy and Zedeck (1983),

"There is little disagreement that if done well, performance measurements and feedback can play a valuable role in effecting the grand compromise between the needs of the individual and the needs of the organization."

(p. 2)

Sample (1986) suggests that performance evaluation is one of the most important managerial functions. Drucker (1974), Ouchi (1981), Peters and Waterman (1984), Peters and Austin (1985), Peters (1987), and other contemporary writers on the subject of business administration all contend that successful organizations of the future will be those which do the best job of managing their human resources and that human resources are and will be the most important assets an organization can have. To be valuable, human performance must be guided and controlled but the degree to which anything can be guided and controlled is a direct function of the accuracy with which the performance can be measured. It would seem self evident that performance evaluation deserves a great deal more attention than it receives in some organizations.

Purpose

Steers and Lee (1983) suggest that there are three goals which a performance appraisal system should strive to attain. They are: (1) evaluating employees contribution towards the accomplishment of organizational goals, (2) the guidance and development of individuals within the enterprise, and (3) motivation -- "to energize, direct, and sustain behavior." These goals must
be met through a variety of managerial functions and decisions. There are five purposes of performance evaluation generally agreed upon (Sample, 1986; Jacobs, 1986; Guion, 1986; Steers and Lee, 1983) and which require the type of data a good assessment program should yield. These are: (1) selection and placement, (2) training and development, (3) performance feedback, (4) reward and discipline, and (5) organizational development. All of these functions require both behavioral and outcome-based data and all involve important decisions which can have costly consequences for the organization.

People have always been interested in assessing human performance. From the earliest times judgements have been made about performance just as they continue to be made today. According to Mohrman and Lawler (1983), "we are constantly assessing the behavior of ourselves and others both privately and publicly." In the past our judgments may have been more objective since they were usually based on what Gilbert (1979) would call "worthy accomplishments", the amount of fish caught, buffalo killed, land cleared, etc.. Such measures were very concrete and relatively easy to make. Today, in a majority of cases our measures are relatively unsystematic and unscientific or as Borman (1983) says, we frequently make subjective evaluations of a person and then look for evidence to support the evaluation. Nevertheless, the fact is that we cannot escape making judgements about other's performance or having judgements made about our own. Therefore, the pertinent question is not whether we can measure human performance but how well can we do it?
Methods

What is the best way to measure human performance? Performance evaluation methods can be categorized as either direct or indirect. Direct methods measure a product or measureable achievement which is entirely external to the person responsible for it. Indirect methods attempt to measure the behaviour or other characteristics of the person which are thought to contribute to the achievement of a purpose which supposedly cannot be accurately assessed. Indirect methods are of two types: (1) behaviour based, and (2) attribute based. Behaviour based evaluations are probably the most common and, when properly grounded in a thorough job analysis, have been found to be acceptable and legally defensible in the courts. Attribute based evaluations, although showing face validity and high functionality, are very subjective and arbitrary. Attributes are personal characteristics such as intelligence, integrity, empathy, et cetera, which people are assumed to possess and which it is assumed lead to positive performance. Research shows poor correlations between human attributes and particular job functions, thus, these evaluations have been found to be indefensible in court and most experts advise against their use.

The argument over the advantages and disadvantages of behaviour-based versus achievement-based evaluations is a major source of contention between experts in the field of performance evaluation. Gilbert (1979) advocates the assessment of performance on the basis of "worthy accomplishment" because behaviour by itself does not necessarily result in a worthwhile achievement. We must also consider the desired end and the direction in which the behaviour is oriented. It is clear that Gilbert goes beyond the measurement of a product or a narrow range of objectives and in
so doing answers many of the objections raised by detractors of "outcomes-based evaluations." Staw (1983) objects to measuring outcomes on the basis that,

"Much of the difficulty in assessing individual performance comes from trying to measure outcomes rather than behaviors. For many jobs, outcome measures are not readily available and, in some cases, the specification of outcomes for evaluation is counterproductive." (p. 35) "Doing well at the wrong thing sometimes only compounds the problem ...." (p. 36)

Gilbert would certainly agree with the last part of the statement.

Not only may concentrating on behaviour without considering overall organizational goals lead to excellent but misdirected behavior, there is usually more than one way in which a given end may be achieved. According to Gilbert (1979), "the failure to plan the direction of performance promotes incompetence" and "the regulation of behavior promotes incompetence." Behaviour that is too closely regulated negates any possibility of innovative action and innovative action by employees is one of the hallmarks of successful organizations according to Tom Peters (1984, 1985, and 1987).

Bernardin (1986) states, "There are few jobs in which countable results could adequately represent the entire domain of job performance"; therefore, he advocates the use of hybrid methods of performance evaluation. Since even Gilbert (1979) admits that measures of outcomes often fail to provide the guidance necessary to improve performance, a system utilizing multiple performance indicators is probably superior. (Berk, 1986) Assessment centres are programs which use situational exercises and job simulations to "describe, to evaluate, and to predict management effectiveness". (Byham and Thornton, 1986) Such programs use multiple techniques to assess individuals for selection and promotional purposes where the cost of making the wrong decision is
considered to be greater than the relatively high cost of a valid and objective assessment.

Most personnel evaluations are carried out on a periodic basis; that is, they are an exercise in which a supervisor rates his subordinates once or perhaps twice a year and then calls them into the office for an interview. The supervisor may or may not have much knowledge of the job being performed by the subordinate but it will be a rare case where the supervisor will have any knowledge or training in the assessment of human performance. This is an important reason why most performance evaluations fall far short of what they could be and in some cases are worse than no appraisal at all. In fact Guion (1986) is convinced that rater competence is as important as the rating format used and rater training is a major requirement to improve the reliability of assessments. Bernardin (1986) categorizes rating formats as standardized or individualized. Standardized formats are used where the organizational emphasis is on data to be used for personnel decisions such as promotion, retention, or discipline; whereas individualized formats are more valuable in providing feedback on performance for individual improvement, training, and development.

The instruments employed in a performance evaluation are very important in establishing the validity, reliability, and objectivity of these assessments. Many writers (Berk, 1986; Fine, 1986; Jacobs, 1986; Borman, 1986; Siegel, 1986; Byham and Thornton, 1986; Cascio, 1986; Sokol and Oresick, 1986; Rosinger et al, 1982; Nathan and Cascio, 1986) appear to agree that the foundation of any good assessment is a thorough and comprehensive job analysis and this is in fact also the basis of any legally defensible assessment scheme. Sample (1986) believes there are basically two problems to be overcome in implementing a new evaluation system. The first is the development of relevant, legally defensible criteria
on which to rate and the second is to gain acceptance of the system by both raters and ratees. His solution is Behaviourally Anchored Rating Scales (BARS) which are developed jointly by raters and ratees. Such joint development and decision making have a better chance of gaining the support and commitment of those being rated and those doing the rating. (Steers and Lee, 1983)

One of the principal ways in which to ensure employee acceptance of a performance evaluation scheme is to involve them in a thorough and comprehensive job analysis from which the specific job dimensions and the performance criteria are derived. The focus of the job analysis is twofold: (1) what gets done -- the accomplishments, and (2) what the worker does -- behaviour. From this information an assessment instrument is developed which is designed to measure a limited number of factors for which very specific behavioural criteria (anchors) have been determined. The behavioural criteria which will form the anchors for BARS are usually developed using a critical incident approach in which supervisors and employees are asked to identify and describe a large number of incidents which represent unacceptable to exceptional behaviour. The sensitivity of the instrument, that is, its ability to discriminate effectively between poor, average, and exemplary performance is largely dependent on the accuracy with which the behavioural criteria are described. Although BARS may be developed using from three to seven anchor points, it should be pointed out that even a very good instrument is seldom effective in making more than the above noted three discriminations. Once a scale has been developed for each job function to be rated; when comprehensive and systematic procedures have been developed for the assessment; and the raters have been thoroughly trained in the principles and process of performance evaluation, then the system is ready for implementation.
Landy and Zedeck (1983) note four stages of a human performance rating system. They are: (1) observation, (2) storage, (3) retrieval, and (4) judgment. Observations can be done on either a systematic or a spontaneous basis. In most cases spontaneous observations are all that are used in government, business and industrial settings. The problem with this is that exceptional behaviour, either good or bad, is more likely to be noted than average behaviour. This is compounded by the fact that most storage and retrieval is simply a function of human memory instead of an accurate and reliable recording system. The retrieval typically occurs on that once a year occasion when the supervisor is obliged to fill in the performance evaluation report on the subordinate. At that time the supervisor has to try and recall incidents that indicate the behaviour of the worker and human memory being as imperfect as it is, those incidents that were out of the ordinary are more likely to be remembered. The resulting judgements often include unintentional and possibly intentional personal bias in addition to types of statistical bias which often enter into performance evaluations conducted by untrained evaluators.

A good performance evaluation system will provide an accurate description of the ratees performance in a particular job. It provides information useful for making decisions regarding rewards or disciplinary action, performance feedback, or training and development. It will not provide valid information for decision making related to promotion or future placement unless the future position is very similar to the old one. Research shows that past performance is only a good predictor of future performance where the situations and performances required are similar. Since promotion often involves a move into a position with very different functions and requirements, the need for some method of assessing
human potential with a greater predictive validity has led to the development of assessment centres.

**Assessment Centres**

Assessment centres, or programmes as they are referred to in Britain, are not places; rather, they are sophisticated tools designed to assist management in the selection of personnel for higher level or supervisory positions.

**Rationale**

The higher in the organizational hierarchy a position is, the more critical the decisions and their consequences for the organization. Thus the importance of good selection, promotional, and placement decisions should not be underestimated. Stewart and Stewart (1981) state that, "...the hiring of a manager or a professional ... is almost the longest-lasting investment any company is likely to make ...".

Although many methods from employment records and interviews to complex psychological testing are used to provide data for making these decisions, most suffer from a lack of validity, specifically predictive validity. Many studies have shown that successful performance in a given area is positively correlated to future performance in a similar function but not necessarily to a function or job which differs from the successful field. For example, high school marks are positively correlated to performance in first year university and undergraduate marks show a good positive correlation with marks in graduate school but, unfortunately studies do not show a good correlation between university marks and later success in professional careers.
Similarly, performance in a given job is positively correlated with performance in other similar positions but not with performance in a related supervisory position. Therefore, an employment record or performance assessment may be useful in placing a person in a very similar situation and may be useless in selecting or promoting people to higher level positions.

Employers typically fail to recognize that different levels of supervision and management may require entirely different types of skills and present radically different challenges. Poor performance at any given job may be due to lack of human capacity but it can also be due to a lack of challenge and incentive to perform as could be the case when a person's capabilities are not sufficiently challenged or rewarded. Such a person may perform to a much higher level in a more challenging position but in most cases they will never get the opportunity unless they leave and go to another organization. Stewart and Stewart (1981) say that assessment programs will find this potential and in addition by providing assessees with a taste of the higher level position, also permit potential candidates to decide whether or not they really wish to advance to another level. In this way the company may find good managers they would otherwise overlook or it may save them from promoting an otherwise good employee to his level of incompetence. Either mistake can be costly!

**Description**

An assessment program is a procedure in which a small group of candidates is put through a series of exercises and tests which are designed to assess their ability to perform in a job other than the one in which they have been engaged. These programs may run from one day to two weeks with three days being the most typical duration. A wide variety of exercises may be used
depending on the particular abilities to be assessed. Although "canned" assessment programs are available, Stewart and Stewart (1981) caution that since each organization is different in some ways from every other one, a valid assessment programme needs to be designed for each particular situation.

The first step is to determine the criteria which are accepted by the particular organization as being essential for successful performance in that organization at the level under consideration. Once these criteria have been derived, then exercises and tests must be selected or designed which will permit trained observers to differentiate between good and poor performance on the criteria. Observers must be selected, trained, and tested to obtain a high degree of rater reliability. A competent administrator is essential to ensure an efficient and smoothly operating process which does not detrimentally effect the performance of any of the candidates. The candidates are selected, the programme is executed, and the results utilized in the decision making process.

Stewart and Stewart (1981) make it quite clear that there is no "ideal type" of manager for which generic criteria could be derived which would then be applicable across a wide range of situations. This seems to me to be the simple recognition that human performance is a function of the environment in which it occurs, as well as of human capacities. (Gilbert, 1979, 1982a, 1982b) Because of the wide variation in corporate environments, the types of business involved, the nature of the corporate hierarchy, the communication patterns; in short, all those factors which have been amply identified by Drucker (1974), Peters (1984, 1985, 1987), and others as making up the corporate "climate", it is essential to derive criteria of effectiveness which matches the particular organization. In addition, criteria which are derived from consultations with people in the organization are more likely to be acceptable than those which could be imposed from an external
source. As with any performance assessment scheme, acceptability is a key component of a successful system. Unless thorough job analysis data already exists the first step in deriving the criteria is to conduct such a job analysis to ensure the validity of the data.

There are a wide variety of exercises and tests which can be used in an assessment program including group or "task force" exercises and simulation games designed to assess ability to work with others in problem solving situations. Individual exercises include writing exercises, presentations, and the "in-basket" test where the individual is presented with a large number of tasks and required to prioritize them and indicate how they would deal with them. Time and data, or lack of them, are always factors in the in-basket tests. Psychological tests may also be used to assess individual traits.

The observers selected may be permanent professional staff or may be selected from line managers at least one step above the positions for which the persons are being assessed. Permanent observers are typically used in Britain on civil service selection boards while line managers are usually utilized in the North American context. Either method works well provided that the criteria are limited in number, explicitly stated, and the observers are well trained to ensure inter-rater reliability. The use of different line managers for different assessment programs does involve more observer training but has the added advantage of being an excellent training and development exercise in the ongoing assessment of human performance.

The basic problems of objectivity, reliability, and validity are the same for assessment programs as they are for any other measurement and evaluation system. Objectivity and reliability are primarily achieved through well developed observation and recording instruments and highly trained and tested observers.
The observer must be familiar with the basic concepts of an assessment program, the specific criteria utilized, and the administration of the program. They must be able to devote their attention to observing, categorizing, and recording data and not be distracted by administrative details. The judgment or evaluation of data should be done following, not during the sessions to ensure the greatest accuracy of the recorded data.

Stewart and Stewart (1981) say the key question for observers to ask while recording data is, "How do I know that?" If the answer is, "because that is what the participant did", then the data are probably good, whereas, if the answer is, "because he did such and such, from which I deduced that ..." (p. 146), then it is a poor observation. Any person observing human behaviour for the purpose of assessing it, whether manager or instructor, would be well advised to carefully consider this question and their own answers. As Stewart and Stewart (1981) point out, observers must "... concentrate on recording participants behavior and not their own feelings about their reactions to the participant." (p. 146)

Validity

Validity of assessment programs is properly considered under five different categories. These programs are commonly considered to have very high face validity. That is, they are generally seen both by management and by assessees as being very valid for their purpose, thus, the public relations value of these programs should not be underestimated. A properly designed assessment program should be very high on content validity which means the tasks selected are a good simulation of the tasks which an incumbent of the position(s) under consideration would have to perform and provide adequate coverage of the full range of tasks for which an incumbent would be responsible.
Concurrent validity is a measure of the degree of success present managers would have in the assessment program. This is determined by putting present incumbents through the program and measuring the outcomes. Stewart and Stewart (1981) claim high concurrent validity for well developed programs. Concurrent validity is not a substitute for predictive validity but without it predictive validity would necessarily be doubtful since we would have to conclude either that present managers are very poor and the criteria adopted were a great improvement, or that the criteria selected were not relevant.

Construct validity refers to the question of whether or not management potential is actually being assessed? Satisfactory construct validity is claimed on much the same basis as predictive validity. For predictive validity it must be shown that participants who score high on the assessment program are also evaluated positively in their future performance at a higher level of management while those who do less well on the assessment program also perform less well in their future management assignments. The results of the assessment program should enable us to accurately predict who the best managers will and will not be and, therefore, who should be selected or promoted. Stewart and Stewart (1981) cite many different studies, most of which appear to support their argument that predictive validity does exist in good assessment programs.

Disadvantages

In Britain assessment programs have been utilized primarily for the selection and promotion of civil servants and management personnel while in the United States much more emphasis has been placed on the value of such programs for training and development.
In this context participants are given counselling and feedback on their performance and may even be offered assistance in developing plans to rectify any perceived deficiencies. There is no doubt that assessment programs are more expensive to conduct than most other methods used for selection and promotional processes and only the promise of more accurate and thus less costly decision making can justify the whole procedure unless it is also used for training and development. Although high cost in both time and resources is undoubtedly the greatest single disadvantage of assessment centres, another consideration is the strong tendency of such programs to perpetuate the existing system and exclude those persons who could bring new and innovative ideas to the management structure. Studies have shown that people who could be considered to be low on conformance and/or high on independence are often excluded early in most selection processes including assessment programs. Arguments can be made both in favour of and against the desirability of this exclusion, however, many contemporary management theorists (Peters, Waterman, Austin, Drucker, etc.) feel strongly that excluding this type of talent is very detrimental to the competitive positioning of any organization in a world of rapidly increasing change and competition.

**Educational Measurement**

Assessment programs are the epitome of performance assessment in the working world and it is at this point where the assessment of performance in industry is most closely related to performance assessment in educational institutions. Both domains have a similar problem but each has a different advantage. The problem is how to maintain a realistic measurement environment while maintaining enough control to ensure reliable measurement.
An institution has the greatest difficulty in providing the realistic environment and the least difficulty in controlling the measurement situation. The business world has the realistic environment as a given, but, has difficulty controlling the situation well enough to provide reliability of measurement. The assessment centre which uses realistic simulations in controlled conditions with well trained raters seems to be the best solution to the problem.

Educational institutions have the same five requirements in their measurement systems as do business and industry: (1) validity, (2) reliability, (3) objectivity, (4) sensitivity, and (5) functionality. However, the necessity of attaining these requirements may be recognized to a greater degree in education than in industry since educational measurement has evolved into another discipline. Educators have a long history of concern with the ability to measure aptitude, achievement, intelligence, and other characteristics of humans. In the process, two systems of measurement have evolved; norm-referenced measurement and criterion-referenced measurement.

**Norm-referenced and Criterion-referenced Testing**

Norm-referenced testing is designed to discriminate between students, that is, to permit comparisons to be made so the relative status of each student in a group can be determined (Block, 1971; Bloom, 1971; Carroll, 1971; Gronlund, 1981, 1982; Popham, 1978). Many statistical procedures have been developed for use with this type of testing; however, these procedures all depend on a wide range of variance in the test scores (Baker, 1974; Linn, 1979a, 1979b; Nitko, 1983; Popham, 1978). For this reason norm-referenced tests are deliberately designed to produce considerable variance in test scores even if this reduces the content validity of the test (Hoffman, 1962; Popham, 1978; Strenio, 1981). Norm-referenced tests
are normally applied to a broad content area which does not permit the accurate description of a particular performance and thus has little value for diagnostic or remedial purposes in instruction.

Criterion-referenced testing is generally applied over a much narrower domain and is designed to accurately assess performance, that is, what a student can actually do. The emphasis here is the absolute status of the individual student in relation to a predetermined standard. Popham (1978) says it is an "evidence orientation" which is designed to describe as clearly as possible what is meant by the examinee's performance. If the instruction has been effective, then this testing strategy will result in very low score variance with the result that the statistical procedures used with norm-referenced testing are quite useless for criterion-referenced test development and testing (Baker, 1974; Linn, 1979a, 1979b; Popham, 1978).

**Essential Characteristics**

Validity and reliability have been discussed previously and have the same meaning for educators as for business people (Blunt, 1986; Gay, 1980; Gronlund, 1981, 1982; Kerlinger, 1986; Linn 1979a, 1979b; Nitko, 1983; Popham, 1978). In education, objectivity in testing often refers to the method of scoring the test with tests composed of selection type items such as multiple-choice being classified as objective while tests requiring supply type answers, especially essay types, are classified as subjective (Gay, 1980; Gronlund, 1981, 1982; Nitko, 1983; Popham, 1978). Critics of multiple-choice tests (Hoffman, 1962; Pottinger, 1979; and Strenio, 1981) say these tests should be called machine or computer scoreable tests since objectivity is not just a function of the scoring of a test. If the test is not valid and reliable it cannot be objective, but, objectivity means more than just reliability and
validity. An objective test is one in which the expected performance and the standards against which it will be judged are determined and made explicit well before the test and preferably at the beginning of instruction (Bloom, 1971; Popham, 1978; Tinkelman, 1971). Students often equate objectivity with fairness.

Sensitivity refers to the ability to make fine discriminations as a result of the test scores. Norm-referenced testing requires the assurance that student A who scored higher than student B actually is better than student B while criterion-referenced testing requires the assurance that decisions to classify students as masters or non-masters are accurate. Thus, a norm-referenced test requires sensitivity across the entire range of possible scores whereas a criterion-referenced test is primarily concerned with sensitivity at the cutoff or decision point. There is always some possibility that a test score is not an accurate measure of the student's knowledge or ability; therefore, a test score is generally considered to be composed of the true score; and an error component (Kerlinger, 1986). The test score should be regarded as having a tolerance or being within a band of scores the width of which is called the standard error of measurement. If this tolerance is wide, then the test will have poor sensitivity so it is necessary to reduce the error measurement to a minimum.

Reliability and validity are obviously prerequisites for a sensitive test.

Functionality is simply a way of saying that it must be practical to administer and score the test. There is always a trade off between the time and resources required for testing and the quality of the testing situation; therefore, careful decisions must be made regarding the objectives to be achieved and the cost of the alternatives. Those making the decisions must consider whether they are measuring the trivial or the important and whether, if
they have limited resources, they should measure many things relatively poorly or a few things relatively well (Klemp, 1979). The purpose of the test should be a major factor in considering what is functional. A valid, reliable, objective, and sensitive, but not functional test would have no value, but, the decision is more often made in favour of functionality at the expense of the other factors.

Types of Testing

Shirley Boes-Neil (1979) categorizes testing techniques in four categories from the most desirable to the least and summarizes the advantages and disadvantages of each. This categorization runs from on-the-job testing to simulations to projects and performances and finally paper-and-pencil tests. Actual performance situations or on-the-job testing is said to be "ideal testing." Schalock (1981), and Grant and Kohli (1979) are in agreement with Boes-Neil while other writers such as Pottinger (1979a, 1979b), Klemp (1979), and Blank (1982) all express their preference for something other than pencil-and-paper testing which is deemed to be the least desirable, albeit, usually the most practical for administrators. All of these writers are critical of the "objective" or multiple-choice type item; they point out that in practical situations people are not called upon to select one answer from four (Pottinger, 1979a, 1979b) and usually the test items only measure recognition and possibly not even recall of knowledge (Klemp, 1979). Goldsmith (1979) reminds us that even the use of more complex written tests often measures only the students ability to write exams rather than realistic performance or even recall of knowledge.
Performance Testing

Jones and Whittaker (1975) use only two categories, 1) performance testing, which includes Boes-Neil's first three categories, and 2) written tests. They also point out, as do Grant and Kohli (1979) and Fitzpatrick and Morrison (1971), that there is no definitive line between performance tests and written tests because in some situations a written test is a performance test. Whether a test is a performance test or not is primarily a matter of the degree of realism in the testing situation.

A performance test "is one in which some criterion situation is simulated to a much greater degree than is represented by the usual paper-and-pencil test." (Fitzpatrick and Morrison, 1971 p. 238)

Performance testing has been badly neglected, especially in education, as a result of the belief in a high relationship between knowledge about a job and actual performance. Jones and Whittaker (1975), cite studies which show otherwise and therefore claim that direct measures are essential, as knowledge is a necessary but not sufficient condition to guarantee ability to perform. They also point out that performance testing is not just the informal employee observation that usually passes for employee evaluation but is, instead, a more structured situation utilizing a set task under specified and controlled conditions with objective evaluation.

Jones and Whittaker (1975) outline the necessity of job analyses, task analysis, the writing of behavioural objectives, and the design of good test specifications (blueprints) as the proper procedure for the creation of valid, reliable, and objective performance measures. They classify testing two dimensionally with one dimension being hands-on and hands-off testing and the other dimension representing fixed-sequence or variable-sequence testing.
These characteristics define the degree of realism or fidelity of the testing situation. On-the-job testing, while having the greatest degree of realism, usually exhibits the lowest degree of control and thus of reliability, objectivity, and sometimes even validity. Thus some form of simulation is usually the best form of testing since a simulation represents a high-fidelity, variable sequence test.

Fidelity, or degree of realism, has a strong influence on the validity of performance testing but fidelity is not established easily. Most performances are not unidimensional and succeeding cases requiring the performance usually do not occur in a single real-life situation but in a large variety of situations. Obtaining a sample of performance which is truly representative of the situational variance and at the same time maintaining enough control to ensure reliability and objectivity is a major challenge in the development of tests to measure human performance.

Performance testing assesses either the product, the processes, or both and is accomplished through structured observations of the examinees. The most common instruments used are checklists and rating scales designed to assess quality, quantity, process, speed, or a combination of these (Fitzpatrick and Morrison, 1971; Jones and Whittaker, 1975).

**Written Tests**

Although Jones and Whittaker (1975) believe that performance testing is essential to determine what people can actually do, they also indicate that written tests serve a valuable function to "round out" the assessment process. The literature on written or knowledge achievement tests is voluminous and goes back many years, however, there is general agreement among most writers on the desirable characteristics of such tests and the methods of construction. Gay (1980), Gronlund (1981, 1982), Henrysson (1971),
Jones and Whittaker (1975), Nitko (1983), Tinkelman (1971), Wesman (1971), and others all seem to agree on these matters. Most writers including Bloom (1956) also seem to believe that it is possible to write objective multiple-choice items that can measure educational outcomes beyond the knowledge level of the cognitive taxonomy. They do not say that it is easy; in fact, time and again one comes across admonitions to the effect that writing good items from well developed test specifications is a very demanding intellectual exercise (Gronlund, 1981; Pottinger, 1979; Wesman, 1971).

The development of statistical procedures for the improvement of multiple-choice items is now well advanced and concepts of item difficulty, discriminating power, etc are well understood and accepted as they apply to traditional norm-referenced testing. Gronlund (1981), Jones and Whittaker (1975), Linn (1979), Nitko (1983), Popham (1979), and others have all shown that these procedures are not valid or useful for analysis of criterion-referenced tests primarily because these tests have low variance. This problem is even worse when we move to individualized testing since traditional statistical procedures are based on group testing and designed to improve a particular group test (Baker, 1974).

**Computer-assisted Testing**

Computers are useful in the construction and administration of tests where at least one of the following conditions holds:

1) tests constructed to local specifications are required.
2) frequent testing is required.
3) multiple forms of each test are required.
4) individualized testing is desirable.

Competency-based education and mastery learning are prime situations since they require one or more of the above conditions. In fact some writers say that computers to administer testing is a
technological necessity to make the concept of mastery learning practical (Dunkleberger and Heikkinen, 1983).

Lippey (1974) lists five functions which can be performed by a computer: (1) item banking, (2) item generation, (3) item attribute banking, (4) item selection, and (5) test printing and then he also mentions three other functions which could be added to the list: (6) test scoring, (7) maintaining records of students, tests, and items, and (8) diagnosis and remediation. Diagnosis and remediation is also mentioned by Stodola (1974) and consists of having the computer recommend other sources to study for the student who selects a wrong answer. This involves a very careful and thorough educational program development and has not often been accomplished.

Baker (1974) reviews the statistical procedures for item analysis and improvement and notes that although the computer is ideally suited to this kind of work, the procedures are not applicable to individualized testing. He also criticizes the common method of generating parallel test forms by random selection from the test bank stating that this is frequently an inadequate method. What is required is a well defined universe of items, not just a collection. All authors (Baker, 1974; Choppin, 1985; Feuer, 1986; Prosser, 1974; Stodola, 1974) appear to agree that an item bank must be more than just a collection of items and should be assembled using well developed test specifications. "Simply selecting items at random from a collection of items does not insure the creation of randomly parallel tests." Baker (1974)

Hsu (1985) concludes that administration of testing by computer is only justifiable if it can improve the quality of testing and he suggests there are four areas in which this is possible: (1) by providing immediate feedback to students, (2) by adaptive testing, which is a system whereby the computer selects the next item on the basis of the response to the last item, (3) by storing
and analyzing test results, and (4) by increasing test security. However, Hsu also notes that he is aware of no studies which show that the quality of assessment is or has been improved by using a computer. Hsu (1985) reviews the various programs cited in the literature and concludes that computer-assisted testing has so far been very limited to simple functions such as item banking and test scoring. He sees great potential in the development of adaptive testing but is not impressed by the present state of the art.

**Performance Assessment as a Performance**

Performance assessment, whether done for educational or business purposes, is a performance and can be analyzed by utilizing the seven element performance model.

**Knowledge**

Starting with the human factor the first requirement is knowledge. The rater must have some knowledge of the job for which the ratee is being rated, (Bernardin, 1986; Ilgen, 1983) that does not mean the rater has to be able to do the job or know all the details of it (Gilbert, 1978; Vineberg and Joyner, 1983). It may only be necessary to know the outcomes and the standards for those outcomes. On the other hand the rater must have knowledge of the process of performance assessment (Gronlund, 1981) and this is the most common deficiency in most raters. It is a common fallacy that if you can do a job yourself, then, you are automatically capable of rating someone else's performance on the same type of job.

Performance assessments can be affected by many different biases of the rater and, although there is no way to eliminate these biases, the best protection seems to be rater awareness of them.
Most writers discuss the occurrence of frequent measurement biases such as halo effect, central tendency effect, or the variable effects of raters who rate everyone high or low (Gronlund, 1981; Guion, 1986; Mead, 1986), and opportunity bias (how often does the rater have the opportunity to observe the ratee) (Bernardin, 1986; Mitchell, 1983; Vineberg and Joyner, 1983). A smaller and more recent body of literature deals with what are called personal biases. These include: racism, sexism, congruency of sex and sex role, (Guion, 1983; Ilgen, 1983) sexual preferences, stereotypes, (O'Leary and Hansen, 1983) physical and mental handicaps, physical attractiveness, (Guion, 1983; O'Leary and Hansen, 1983) and the idea that there is one right way and only one way to do any job (Guion, 1983; Mead, 1986). Other variables which can affect a performance assessment are: the power of the rater, the experience of the rater, the similarity between the rater and the ratee (Ilgen, 1983), the social role of the ratee and whether the ratee is a part of the "in group" (O'Leary and Hansen, 1983), the ratees desire for the positive opinion of others (either coworkers or managers), the managers confidence in the ratee, (O'Leary and Hansen, 1983; Wexley, 1986) the inter-personal skills of the rater and ratee (Guion, 1983), and the frequency of contact and communications between them (Guion, 1983; Mitchell, 1983).

Skills

The performance rater needs skills in observing performance, recording what was observed, classifying, and reporting on the performance. The latter skill is largely concerned with the typical performance assessment interview and the rater should realize that not only the assessment, but, how it is reported can have an influence on the future performance of the ratee (Curran, 1983; Mitchell, 1983). The cognitive complexity of the rater; that is, the
rater's ability to discriminate between different levels of performance and handle the individual differences of people, can impact the value of the performance assessment. (Guion, 1983)

**Motives**

The motivation of the performance rater is one area which is usually totally ignored. Performance assessment is often regarded by some as an odious task which is a requirement but not a central function of the rater and, therefore, is seldom recognized or rewarded. It is an extra task which does not contribute directly to the main function of the job and therefore seldom contributes to the attainment of the personal goals of the rater. Raters who are personally involved in the assessment process, who accept the methodology utilized, and who believe it is a legitimate function of their job, are more likely to produce valid assessments (Steers and Lee, 1983; Mohrman and Lawler, 1983).

**Data**

On the environmental side of the performance assessment model, the organization must provide explicit information to raters about the organizations expectations and how to achieve them (Cascio, 1986; Mitchell, 1983; Schneier, Beatty, and Baird, 1986). Test manuals and instructions to raters including, specific criteria and standards, contribute to inter and intra-rater reliability (Fitzpatrick and Morrison, 1971; Mead, 1986; Siegel, 1986; Quellmalz, 1986). The proper training of raters is one of the best methods of ensuring high quality assessments and inter-rater reliability (Fitzpatrick and Morrison, 1971; Guion, 1986; Quellmalz, 1986; Rosinger et al, 1986; Schneier, Beatty, and Baird, 1986). In addition, although rarely done, the raters must be provided with
feedback on the quality of the assessments they are doing (Byham and Thornton, 1986; Gilbert, 1978; Mitchell, 1983; Sokol and Oresick, 1986; Wexley, 1986).

Instrumentation

Instruments, which usually means the rating formats and/or performance tests used, must be developed from a rigorous job analysis in order to ensure a high degree of validity (Nathan and Cascio, 1986; Fine, 1986; Siegel, 1986; Byham and Thornton, 1986; Rosinger et al, 1983). Performance tests should be as high in fidelity as possible (Fitzpatrick and Morrison, 1971; Jones and Whittaker, 1975; Siegel, 1986; Vineberg and Joyner, 1983) and, as with BARS, the criteria and standards must be explicit and unambiguous (Borman, 1986; Jacobs, 1986; Sample, 1983; Schneier, Beatty, and Baird, 1986; Sokol and Oresick, 1986; Quellmalz, 1986). Whenever possible, multiple methods should be utilized to overcome the limitations imposed by any one method (Bernardin, 1986; Gronlund, 1981; Jacobs, 1986; Quellmalz, 1986; Schneier, Beatty, and Baird, 1986). The instrument should assess outcomes and/or behaviour but not personal attributes. The assessment of personal attributes has not been validated by research, is not legally defensible, and is subject to greater error as a result of assessment biases (Byham and Thornton, 1986; Nathan and Cascio, 1986; Sample, 1983; Sokol and Oresick, 1986; Wexley, 1986).

Incentives

Just as the motivation of the rater is usually overlooked, so too are the incentives the environment offers to the rater for producing good performance assessments. Tom Peters (1984, 1985, 1987) exhorts managers to "reward the behaviour you want to
produce!" The organization must recognize the importance of valid performance assessments to itself as well as to the worker, and reward those who produce valid performance assessments. The rewards can take many forms depending on the circumstances; the point is, if good performance assessments are important, then, make it obvious that they are important. Whatever the rewards, it must also be obvious that they are directly related to the performance assessments being done and not just to the raters overall job performance (Hall, 1983; Ilgen, 1983; Mitchell, 1983; O'Leary and Hansen, 1983).

Climate

The organizational climate also has a great influence on the quality of performance assessments. The organizational structure, both reporting (communications) and authority, influences the quality of performance assessments just as it influences the quality of any other task. Rater commitment to the organization and its goals is largely determined by the climate (Hall, 1983; Mohrman and Lawler, 1983; Schneier, Beatty and Baird, 1986; Steers and Lee, 1983). Managerial support is essential if raters are to take the risks involved in making honest performance assessments.

Application

It can be seen that a thorough knowledge of the performance model and its seven elements coupled with a knowledge of the performance situation to be analyzed will suggest the questions which need to be answered in order to recommend improvements in the system. This is the same approach which Gilbert (1982a, 1982b) recommends under the heading of the PROBE method of performance assessment. PROBE is Gilbert's name for "profiling behaviour." The
approach is valid and useful in examining existing performance systems or designing new ones. It is useful at all levels of performance and at all levels of an organization. It is useful for analyzing the performance of a factory worker or the manager who supervises others, and, as I have shown, it is useful for analyzing performance assessment, which is itself a performance.

Competency-Based Learning

Proponents of competency-based learning or competency-based education tout this system as an alternative to "traditional" education. They make many claims as to the advantages of this system, but, none is more central than that students taught by this method can actually perform useful skills competently.

Definition of Competence

"If competency is to be the primary goal of elementary and secondary education, it is inconceivable that professional educators can continue to practice, not knowing exactly what "competency" means." (Bunda and Sanders, 1979)

"The determination of what constitutes competence is "the rock upon which all defensible examination procedures must be built" (Loveland, 1976/cited by Pottinger, 1979)

"The result of not understanding in the first place what really constitutes competence has resulted in millions of dollars and person years of research directed at predicting outcomes of dubious consequence." (Pottinger, 1979)

Bunda and Sanders, Loveland, and Pottinger have provided us with good reasons for defining competence as accurately and comprehensively as possible. In addition, as Collins (1983, 1986) points out we need to define our terms so that even if we do not agree, at least we have a better chance of understanding the others point of view.
Various dictionaries, Oxford, Websters, and others define competence as: a sufficiency of means for living; the ability to do; having sufficient ability; or, the quality or state of being competent. The word competent is defined as: having sufficient ability; having requisite or adequate ability or qualities; having the capacity to function or develop in a particular way. In addition, competent is suggested as a synonym under the definition of the word sufficient, with the following explanation: competent suggests measuring up to all requirements without question or being adequately adapted to an end.

There are two major problems encountered in trying to agree on a definition of competence. One is that competence is usually seen in one of two ways: as a "hypothetical construct," or as "a measure of performance." The choice of definition has significant consequences for the ways in which we attempt to measure competence (Bunda and Sanders, 1979). The second problem concerns time frames, that is, whether we are interested in past or future performance (Olson and Freeman, 1979). The two problems are related but dissimilar. They are related in that those who see competence as a hypothetical concept are often interested in future activity which cannot be accurately delineated nor the circumstances in which it may occur. Chickering and Claxton (1981) quote Gary Woditsch thus:

"How man chooses to behave today is the best predictor of tomorrow. Now, if what we want tomorrow is yesterday, we can make a good case for defining competence in terms of what is normative in human behavior today ... But if what we want tomorrow is ... some unprecedented combination of yesterday's successes and today's hopes, the way most men behave cannot be our norm. For a humanity that envisions a future better than its past, those capable only of replicating the past must be judged incompetent."

A practical example is the training of astronauts. How are standards set and defined for performances which have never before been accomplished? It may not be as easy as working with
that which is more commonplace but, it can be done to a certain degree. And it is not done by considering the desired end as something hypothetical; it is done by directing attention to the performance which is thought to be necessary to accomplish a specific goal. Therefore, this study defines competence as a judgment made about the value of a measure of human performance.

Collins (1983, 1987) points out the misuse and abuse of the language in regard to the subject of competence and in particular the redundancy of the terms "minimum and maximum competency." If as the dictionaries say competence means a sufficiency or adequacy, then the term implies a minimum and so, if one is competent, then one is at least minimally competent. The term "maximum competency" is not well chosen. Maximum competence implies the best that one can be, exemplary performance in Gilbert's (1979) terms, and might better be replaced with the term "excellence" which Tom Peters (1984, 1985, 1987) uses to refer to exemplary performance in business. If performance is considered as a continuum with zero or no performance at one end and excellence at the other, somewhere along the scale a judgment can be made that at a particular point any performance is competent above the point and incompetent, or in kinder terms, "not yet competent" below the point.

Chickering and Claxton (1979), Klemp (1979), Nickse (1979), and Pottinger (1979b), all agree that competence is both personal and situational. The personal aspects include knowledge, skills or abilities, and motivation while the situational aspects refer to environmental variables. Pottinger (1979b) considers no occupation to be homogeneous, differences exist at different levels of practice and in various situations. He also points out that the domain of performance is dynamic and constantly changing and there are complex interactions between the components -- knowledge, skills, and environment-- which must be considered. Excessive
reductionism, warned against by Chickering and Claxton (1981), Collins (1983, 1984, 1987), and Pottinger (1979, 1979b), the result of looking at the individual components without considering the complex interactions and failing to observe the true nature of human performance. These writers and others agree that competence must be tied to a particular position or role (Chickering and Claxton, 1981; and Schalock, 1981). It is unlikely that the majority of persons will be competent in all roles in all situations, though those who see competence as a hypothetical construct seem to be searching for the "holy grail" that would produce this condition.

Definitions of Competency-Based Learning

There are many different definitions of CBL and CBE and a multiplicity of differing implementations of the concept. Some writers (Blank, 1982) combine a wide range of alternatives under the rubric of CBL. The individual proponents of any given variation are usually emphatically opposed to such a classification pointing out that their instructional format differs in some significant manner from the others. There appear to be two points on which CBL proponents agree: (1) that their system or variation is decidedly superior to all forms of "traditional" education, (Riesman 1979) and (2) that these systems are all performance based; that is, based on the ability to measure the performance of students (Blank, 1982; Klemp, 1979). Traditional training programs are primarily concerned with measuring students' knowledge rather than students' abilities (Blank 1982). McAsham (1979) states that the justification for CBL is its promise of greater learning and student achievement as a result of CBL being a superior educational format.

Grant (1979) defines CBE as,
... a form of education that derives a curriculum from an analysis of a prospective or actual role in modern society and that attempts to certify student progress on the basis of demonstrated performance in some or all aspects of that role. Theoretically, such demonstrations of competence are independent of time served in formal educational settings." (p. 6)

According to Rusnell and Collins (1989),

"The features which define CBE systems are that the competencies are derived from rigorous occupational analyses, verified by practitioners, stated in on-the-job performance terms, known by the trainee in advance of instruction, tested on individual trainees, tested in accordance with industry accepted standards, and each competency is fully mastered by the trainee before the next competency is attempted." (p. 17)

Grant et al (1979) reviewed and analyzed the application of CBE in a number of post secondary liberal arts and professional training settings whereas Rusnell and Collins (1989) reviewed a post secondary vocational training setting at S.I.A.S.T.'s Woodland Campus in Prince Albert. Since the outcomes of trades training are generally seen as being different from those of professional preparation, it is not surprising that the emphasis in a definition would shift somewhat. The emphasis also shifts from the public school system where the emphasis is on the acquisition of basic skills to post-secondary schools where the emphasis is on generic skills related to a particular life role ( Boes-Neill, 1978; Bunda & Sanders, 1979; Goldhammer & Weitzel, 1981; Nickse, 1981).

What is Competency-Based Learning?

According to many of its proponents, competency based learning, although not new, is a non-traditional method of delivering educational programs that when properly implemented is superior to the traditional approach. They emphasize the same points: that competency-based learning is a "learner-centered philosophy"; "goal-based planning is fundamental to the
competency-based education philosophy"; and "expected outcome statements are placed "up front" as guides for both instructors and learners" (Parnell 1978). Having these objectives "up front" is necessary so both the instructor and the student know what is to be learned and both are in a position to judge when it has been learned. It also makes it easier for anyone else, administrators or employers for example, to determine exactly what skills can be expected of students who have successfully completed any given unit of study. With these objectives "up front" and properly stated in behavioural terms, evaluation becomes straightforward since the conditions under which the performance will be evaluated and the required performance criteria are stated in the objective.

Another important characteristic of competency-based learning is that all testing or evaluation is of the "criterion referenced" type and not "norm referenced." Each student is measured against a predetermined standard rather than against other students. For this reason, competency-based learning is often seen to be more "fair" than traditional methods which rely on more "subjective" means of evaluation. A good competency-based program should rely on practical tests requiring performance under realistic conditions rather than just paper and pencil tests. This is based on Blank's two basic principles which underlie the concepts of competency-based learning. "First is the notion that 'human competence' is the ability to actually perform. Knowledge, attitudes, and effort are of little value without results" (Blank, 1982).

The second principle is: "--'mastery learning'-- holds that almost everyone can learn almost anything well if given quality instruction and sufficient time" (Blank, 1982). For this reason time is not emphasized in a CBL program. Since all students need to master the needed skills to at least some minimum level in order to be considered competent and since not all students learn at the same rate, a program that is flexible enough to accommodate these
learner differences is necessary to produce competent performers. Traditional programs often have little room for accommodating individual differences since they are usually locked into a set time frame and a set curriculum which has been devised for the "average" student.

A corollary of this second principle is that not every person learns the same things equally well in the same ways. Therefore, quality instruction as well as sufficient time is required. Quality instruction must therefore be that which takes into consideration individual differences. Some students may learn satisfactorily by reading well or poorly written material, others may not. Some may need to hear and see the instruction at the same time. Although Blank does state that quality instruction is required, most of his book is concerned with the organization of the "systematic approach" (Blank, 1982).

Guskey (1985) on the other hand, emphasizes the learning process. Whereas Blank sees Mastery Learning and CBL as the same thing under a different name, Guskey differentiates between the two on a number of important grounds. He sees Mastery Learning as a group process which can utilize any number, or combination of, teaching techniques. The teacher is the primary source of instruction and the teacher controls the pace of instruction. With CBL the basis of instruction is the individual with the pace of instruction controlled by the individual student. The primary source of instruction is the course materials supplemented by the teacher.

Mastery Learning is also a systematic approach. It involves clearly specified learning objectives and criterion referenced evaluation. However, process is the emphasis. Having determined the objectives and considered student characteristics, the teacher decides on the most appropriate techniques for delivering a relevant learning experience so that the students can attain the
objectives. At regular intervals formative tests are used to assess student progress against the criterion. From the tests the teacher determines what the students difficulties are and prescribes corrective learning activities to help the student overcome them and master the skills and knowledge. These correctives are prescribed on an individual basis for each student and are different from the original instruction.

Following the correctives another formative test which is an alternate form of the first is administered. Students who demonstrate mastery of the subject or skills on the first test do not require the correctives or the second formative test. Instead they are given enrichment activities. When all students have mastered the material to the same high level, the group moves on to learn new material. In many CBL programs, if the student does not master the material to a high level, the student is required to complete exactly the same learning experience completed unsuccessfully. This leads us to the fourth characteristic that differentiates CBL from traditional programs.

In a traditional program all students move on to new material together when the teacher decides the time is appropriate. In many cases some students will not have mastered the material to a very high level and thus they may be handicapped going into the next section of new material. With a CBL or Mastery Learning approach, the student does not move on until the material has been learned to the desired level.

Problems and Criticisms of Competency-Based Learning

Proponents of CBL, such as Blank (1982), have enumerated the many shortcomings of "traditional educational" programs and the impressive gains that are anticipated if a competency-based format is implemented. Collins (1983, 1984, 1987) questions the assumptions
on which CBL programs are founded. Others, (Grant et al, 1979) who have done evaluations of CBL programs find serious problems in their operation but do not dismiss the concept as being ill-conceived.

Collins (1983, 1984, 1987) criticizes CBL for "excessive reductionism" citing the use of long "laundry lists" of insignificant competencies which accompany many CBL programs. Ratcliff (1984) says there is no list of predetermined competencies used in CBL programs and thus no excessive reductionism, she says, however, if there were, then Collins would be right. Zelda Gamson (1979) tells of the programs found during their (Grant et al) evaluations that do develop long lists of specific behaviours and Pottinger (1979) also warns about excessive reductionism in CBL programs.

Collins does not like the behaviouristic foundations of CBL. He derides the behaviourist notion that any action can be broken down into small elements which can then be measured with increased precision while disregarding the motives behind the action. Nickse (cited in Parker, 1984) believes the behaviouristic elements result in improved organization of the educational program and when coupled with humanistic elements, such as a concern for individual learners, results in improved educational programs. Ratcliff (1984) argues that CBL can be accomplished in a humanistic setting while Blank (1982) argues that it is traditional education which is not humanistic. Furthermore, Blank insists that a major advantage of CBL is a systematic approach compared to the traditional approach which is haphazard.

Collins (1984) admits that categorization and some behaviourist elements may be useful as guidelines under some circumstances but not as a formula for developing programs. He agrees that a systematic approach is useful but is adamant that no "overarching" system can solve the wide variety of learning problems encountered in post-secondary education and warns that the danger of having
our attention become fixed on the system as a "panacea" is much greater with CBL. Collins argues that education does not have to be an unsystematic approach and agrees with Rubin (1979) that prefabricated prescriptions are not likely to be of help in solving educational problems. During discussions in class, Professor Collins has made it clear that he is not a defender of the status quo in education and also notes that although CBL has been defined by its adherents, "traditional education" seems only to be defined by default as any form of education which does not use the CBL format.

Gamson (1979) notes that one problem in CBL programs is the lack of "ownership" of programs because they are not developed and delivered by individual instructors, but Blank (1982) says the quality of programs can be improved by making them independent of individual instructors. Rubin (1979) argues that practitioner commitment is essential to the success of CBL programs and Gamson (1979) found CBL programs require more development time, more resources, and especially well qualified people, fully committed, able and willing to do the extra work involved in implementing a CBL program. As she says, "...implementation of an innovation eventually rests with the individual faculty member." (Gamson, 1979)

Ratcliff (1984) cites the variety of approaches to implementing CBL as evidence that centralized control is not desirable because individual innovations are necessary for program improvement. Collins (1983) says that this controlling and restricting influence is a consequence of "... a purposeful attempt to define the teaching-learning situation according to a deterministic doctrine ..." (p. 178) and that what is actually required is professional educators who can design solutions to the myriad variety of problems inherent in post-secondary educations. We need people who can, "... make intelligent plans and modify them according to the circumstances"
Collins claims that CBL leads to increased bureaucratization and centralization of educational decisions and is attractive to administrators because it can provide them with a greater degree of control.

Collins does not like the CBL adoption of industrial-commercial approaches to productivity and efficiency, specifically the reduction of programs to very specific, standardized units, not necessarily appropriate to education (Collins, 1983, 1987). Nickse counters with the argument that, "There is nothing morally wrong with striving for a productive and efficient education system" (cited by Parker, 1984 p. 108). Ratcliff (1984) does not agree that, "all attempts at unitization and standardization are bad" (p.113), but Collins insists, "The proper emphasis for educators resides in the acts of individuals" (Collins, 1987 p. 41).

Goal based planning is fundamental to CBL and the complexity of humans or the situations in which they find themselves ought not to be underestimated; but, it does help to narrow the focus of education (Parnell, 1978). Parnell also believes public education lacks a specific focus, a point which Collins also makes in quoting David Williams who talks about "... public education's proclivity for accepting responsibility for everything. ..." (Collins, 1987 p.39)

Parker (1984) sees CBL as relevant since it focuses on the solving of real-life problems with the process being more important than the actual problems. Collins, however, sees CBL as reinforcing a mechanistic and stereotypical approach to problems when what is needed in a rapidly changing society is "to educate people for intelligent flexibility" and creative thinking (Collins, 1987 p. 41).

Blank (1982) argues that CBL is more efficient than traditional education since it does not waste time teaching irrelevant and nice-to-know material but emphasizes ability to actually perform. The public school curriculum includes a large amount of irrelevant and nice-to-know material which students are compelled to learn. On
the other hand some vocational schools teach such a narrow curriculum that students are ill prepared for any but a very narrow range of highly specific jobs with little or no future. To change jobs or advance to higher positions from such narrow training often entails complete retraining so again it is not a matter of one or the other; it is a matter of making intelligent choices.

CBL is much more demanding of students, requiring those high in personal initiative. A major problem is students who are unable or unwilling to pace themselves with the resulting slow student progress meaning low credit generation, a problem Gamson (1979) says no school can afford to ignore. Unprepared students result in time-lengthened courses rather than time-shortened courses which CBL proponents say is an advantage of CBL. Gamson (1979) says there is never enough time and attention for students in CBL programs and this is no different from traditional education.

"When carefully developed and implemented, the competency-based approach to training is generally superior to the traditional approach in terms of student outcomes and in several other important ways" (Blank, 1982). Blank goes on to say that many studies in a variety of settings show the superiority of CBL and warns that in those which show no difference the individualized approach was usually not carefully designed and implemented. In his well known text, "Foundations of Behavioural Research," Kerlinger (1986) cites a study by Thompson (1980) as being an excellent example of a carefully designed and executed study. This study titled, "Do Individualized Mastery and Traditional Instruction Systems Yield Different Course Effects in College Calculus?", found no significant difference between a well implemented traditional approach and a well implemented Mastery approach. Gamson (1979) notes that many have failed in their attempts to implement the CBL format.

I agree with Collins (1983, 1984, 1987), traditional education
does not have to be unsystematic or to embody the many shortcomings its detractors see in it. The fact that these deficiencies are found in many educational programs is only evidence they can occur. Similarly, I do not believe the faults which Collins finds in CBL are inherent in the concept; they are manifestations of the ways in which particular programs were implemented. The key is in the implementation. No particular format or concept can guarantee any specific results; therefore, the answer is implementation by reasonable, knowledgeable, professional practitioners rather than slavish adherence to a particular doctrine or formula.

Schalock (1981) considers the Achille's heel of CBL to be the quality of assessment. Since CBL emphasizes the "ability to actually perform" (Blank, 1982), the ability to measure actual performance is essential and this ability is not often found in typical paper and pencil testing. Instead, performance testing is required with all the additional development and resources which this entails.

**Measurement of Competence**

Pottinger and Goldsmith (1979) claim that there is general agreement that competence can be measured and Klemp (1979) assures us this is so. If competence represents a judgment point on a performance continuum, then the measurement of competence involves two problems: the first is the measurement of human performance and the second is the establishment of standards on which to base the judgment. The first and most important step is to define the domain of the performance and there does not seem to be any disagreement that the best way to do this is through a thorough and comprehensive job analysis (Pottinger, 1979a, 1979b). The information provided by a systematic job analysis is an
essential minimum in establishing the validity of any performance measure and the validity of the performance standards (Linn, 1979b; Pottinger, 1979a).

Thus, we see the similarity between the problem of assessing human performance in business and in a technical training environment. The problem of establishing an assessment scheme which is valid, reliable, objective, sensitive, and functional is common to all environments and the methods of achieving this goal vary only in detail.
Chapter 3. Design and Methodology

The objective of this study was to test the model by applying it in one practical situation. The model could have been used in many different situations, however, the relatively new Woodland Campus provided the opportunity to do a needed study which promised to be of practical value to the institution. Since Woodland's programs utilize the competency-based format, the ability to make high quality performance assessments is paramount. The computerized testing system provided another unique element.

Process and Product Evaluation

The administration, faculty, and support staff at Woodland Campus purports to train people, measure their performance, and certify their graduates as competent entry level practitioners. Competence has been defined as a measure of performance and performance as the consequence of human behavior interacting with an appropriate environment. In effect, Woodland establishes its environment as a given and assesses its students' behaviours in relation to that environment. The physical facilities of this environment are modern, well equipped and considered to be representative of a modern work environment. If I were following Gilbert's (1978, 1982a, 1982b) advice I would first want to assess the quality of the product; or, as he puts it, "the worthy achievement" of this institution. In this case the products are the students who have been trained and certified by the institution and perhaps the best measure of their worth would be their degree of employability and the expressed satisfaction of their employers with their performance in the workplace.

Unfortunately, the data do not yet exist to make such a judgement. Saskatchewan Education publishes data collected from a
questionnaire sent to all S.I.A.S.T. graduates each year but the return rate for Woodland is low at about 50 percent. There is no subsequent follow up in succeeding years and employers are not consulted so the data are insufficient and incomplete. Some instructors at Woodland maintain contact with some of their former students and with some of the employers of former students but this is a very informal system which is not documented and thus has little or no value for a formal evaluation. Therefore, an assessment of the training process and its immediate outcomes is the only possible basis for evaluation at this time.

Process Evaluation Using the Performance Model

Assessing the process would be the normal second step in any comprehensive evaluation with the purpose of suggesting improvements to the system. I have assessed Woodland’s student evaluation system from the perspective of the seven element model of performance previously described. To produce competent graduates Woodland needs to be able to affect the three personal elements of human performance, namely, the students’ knowledge, skills, and motives; and, to do this within an environment which is as realistic as possible. Woodland must have the ability to measure these elements and also the interactions or consequences of the interactions of these elements with each other and with the campus environment in order to truly be able to assess, judge, and certify human performance as competent.

For this reason data have been collected to answer questions about each of the elements of the performance model as it relates to the student assessment system at Woodland. The study considers the data provided to students for direction and feedback when they are evaluated, the facilities, instruments, and procedures utilized in measuring the various aspects of student behaviour, and
any incentives offered to produce or improve student performance. Since instructors are instrumental in the student evaluations, it was necessary to seek data on the level of knowledge, experience, skill, and motives of the instructors and on the environmental elements which effect the performance of instructors, such as the data provided to them, the facilities and the incentives offered to them.

A Multi-Method Approach

To gather the data required to answer all of these questions a multi-method approach was utilized. Wherever possible, objective data were collected and analyzed using appropriate statistical measures; however, in some cases qualitative data from observations, impressions, and unstructured interviews were gathered where practical and appropriate to obtain.

Assessing the Computer Item Bank

Rationale.

Popham (1978) places the test item improvement processes in two categories: the a priori and the a posteriori. The a priori, or before the fact, method consists of two parts: assessing content validity and assessing technical or mechanical construction of the test items. He describes a method whereby a panel of content experts is used to determine the content validity of an item by comparing it to course objectives, test specifications, and their own knowledge of the particular job or jobs for which the training is designed. This procedure is similar to the way in which an advisory panel might be used to validate the original job analysis and course objectives. The second part of this analytic method is
to have the test items analyzed for proper mechanical construction by someone knowledgeable in this field.

The a posteriori, or after the fact, method is an empirical method that relies on the generation of statistics about each item after the item has been utilized in a test. In other literature this method is often referred to as the "classical item improvement model" and four measures are suggested. The first of them is the "difficulty level" which is a measure of the percentage of students answering this item correctly and not necessarily a measure of difficulty in terms of the level of cognitive processing required to answer the item, or the degree of obscurity of the data requested.

Item discriminating power is a second index which is suggested for evaluating test items and, although there are several different methods suggested for calculating this index, it is basically a matter of determining the percentage of high scoring students who answered the item correctly versus the percentage of low scoring students who answered the item correctly. The idea is that, if the high scoring students answer the item correctly and low scoring students answer it incorrectly, then the item is useful for discriminating between students who know the material or have met the objectives and those who have not.

A third measure is the effectiveness of the distractors which may be calculated in a similar manner to the item discriminating power only now it is done for each distractor in each item. If the distractors have all been constructed so they are all plausible to the uninformed, then each distractor should attract an equal proportion of incorrect responses. If few if any students choose a given distractor, then it probably is not a plausible answer to the uninformed student; as such it is not contributing to the purpose of distractors and should be changed.

The fourth measure suggested is "sensitivity to instruction" which is a measure of how well students answer the item after
instruction compared to before the instruction. This can only be determined by using a pretest/post-test method and, although this method is said to be particularly relevant to criterion-referenced testing and mastery learning situations, it is not always practical since it involves two administrations of the same test.

It can be seen that the empirical "a posteriori" method has serious deficiencies for our purposes. In the first place, these measures were developed for use with norm-referenced testing with the wide variance in scores that is usually an essential characteristic of this type of test and, therefore, are poorly suited for use in criterion-referenced testing where score variance is usually at a minimum. For example, in norm-referenced testing a high positive discrimination index is desirable and a low positive or negative index is highly undesirable while in criterion-referenced testing even a low positive index may be desirable since it may be evidence that the outcome of concern was adequately covered in the instruction.

Secondly, it can be seen that these methods rely on the administration of tests to groups of people, preferably large groups, and thus are not useful measures where individualized testing is used, especially individualized testing done by having a computer capable of generating a different test for each student. Baker (1974) concluded that at best only item difficulty could really be calculated from the aggregation of individual responses which is the case in this study. Determining the percentage of students who have answered an item correctly does not tell us whether the item discriminates between good and poor students and it certainly doesn't tell us anything about the quality of the test item. The index of discriminating power, which would supply a little more information, is not usable because not only does it require administration to a large group, but it is also related only to one test and not to multiple related tests as generated by a computer.
Therefore, the empirical "a posteriori" approach to item analysis and improvement is not applicable to the testing situation in use at Woodland campus.

Many writers recognize the need for better empirical methods to assist in item improvement for individualized and criterion-referenced testing but so far the "state of the art" is not well developed in this regard and even staunch supporters of the "classical" approach are forced to admit the very limited usefulness of their methods in these non norm-referenced situations. Popham (1978), predicted that the "a priori" method would become the most popular method in the coming years. This method is certainly a practical method which yields more information about test items and can be applied before the administration of any tests. This method is based on highly specific test specifications and systematic human judgement. Practicing instructors and teachers can be taught to apply this method without any training or background in statistics, however, this does not mean it is a simple mechanistic system, and in fact, Baker (1974), Gronlund (1981), and Popham (1978) all emphasize that the writing of good multiple-choice test items is a very demanding intellectual exercise.

Method.

Since most knowledge testing at Woodland is done with a large computer testing system which utilizes a bank of approximately 73,000 test items, the computer was used to generate a random sample of 250 items from this test bank. Most of these items (240) were multiple-choice items with the few remaining being binary choice (true and false) or fill in the blank. The multiple-choice items were rated against twelve criteria developed from the literature which were related to the construction of high quality
items. These criteria were derived from Gay (1980), Gronlund (1981) (1982), Jones and Whittaker (1975), and Nitko (1983).

(1) A single clearly formulated problem is presented in the stem of the question. The problem may be presented either as a question or as an incomplete statement but in either case a knowledgeable person could answer it without looking at the alternatives offered.

(2) As much of the wording in the question as possible is included in the stem. There should not be any repetitious wording at the beginning of each of the alternatives.

(3) Clear, concise, simple and unambiguous language ought to be used without irrelevant material or difficult words or unnecessary technical terms.

(4) The stem is worded in positive terms or, if the negative is used, the negative terms are emphasized. Examples: NO, not, never, etc.).

(5) The item ought not to use any specific determiners such as never, always, usually, sometimes.

(6) Distractors should be plausible and attractive to the uninformed as well as homogeneous in type.

(7) Alternative answers ought to be grammatically consistent with the stem and parallel in form.

(8) There ought not to be any similarity of wording between the stem and the correct answer which could provide a clue to the correct answer.

(9) The alternatives ought not to overlap, be all inclusive, and two or more ought not to have the same meaning.

(10) The correct answer ought not to be stated in greater detail or be longer than the other alternatives.
(11) "All of the above" or "None of the above" ought not to be used as alternatives.

(12) Questions should measure higher levels of cognitive knowledge than just recall or recognition of facts, or principles. In other words, questions ought to determine whether the student understands a concept, is able to apply a principle, or analyze a problem. (Higher than objectives for Bloom's taxonomy level 1)

Unfortunately, two important questions which ought to be asked in a study such as this, could not be asked. They are: (1) Does each item measure an important aspect of the occupation and, (2) Does each distractor represent a common student error or misunderstanding? It would have required a person or persons knowledgeable in each trade area represented by the sample to answer these questions, or possibly an extensive computer analysis and these were not possible for this study.

A group of vocational education students from the University of Saskatchewan were used to establish a check on the ratings by the primary rater. This was accomplished by dividing the sample of items into sub-samples which were then rated by the students using the same criteria as the primary rater. These students were nearing completion of a class on testing and measurement and were familiar with the literature and criteria used to rate the items. All students rated a common sample which was chosen to represent what the primary rater considered perfect items and items containing each of the types of errors represented by the evaluation criteria. Pairs of students then rated randomly selected sub-samples. The students ratings were compared with each other and with the primary rater's ratings. A percentage of agreement was calculated and all the percentages were averaged. A high level
of agreement (80 percent or higher) would indicate that persons with similar training would judge the items similarly, and thus the ratings of the primary rater were reliable. In addition, it would indicate that, if Woodland’s staff were similarly trained, they could use the training to improve the quality of the test bank.

Assessing the Computer Generated Tests

A good item bank is only the first step towards the creation of a good testing system. It is also necessary to be able to generate high quality multiple versions of each test in order to satisfy the logistics of such a system. To assess this the computer was used to generate a random sample of 50 knowledge tests and then another 50 which were equivalent to the first 50. These tests were rated against eight criteria developed from the literature on the characteristics of good student tests. The dual versions were rated against each other as a test of equivalency. These criteria were derived from Gay (1980), Gronlund (1981)(1982), Jones and Whittaker (1975), and Nitko (1983).

1. Are there sufficient questions to establish reliability and validity? (10 to 20, never less than 10)
2. Is every item independent of all other items?
   (One item does NOT give away the answer to another or is not just a reworded version of another)
3. Are the questions presented in a logical order such as increasing difficulty, or grouped by objective?
4. Does the position of the correct answer vary?
5. Are the alternatives in a logical order, if one exists?
6. Does the relative length of the correct answer vary?
(7) Are there any questions which would test higher levels of learning than just knowledge?

(8) Is each item related to the stated learning objective?

A check on the ratings of the primary rater was established in a similar manner to that previously described for the item analysis. The same vocational education students were randomly provided with sub-samples of the computer generated tests and asked to rate them. These were then compared with the ratings of the primary rater to provide a measure of agreement.

Assessing the Skills Tests

The most important requirement of a competency-based vocational education program is the ability to measure "actual performance" as opposed to some substitute via a paper-and-pencil test. All proponents of competency-based learning base their arguments for the superiority of this instructional strategy on the ability of graduates to actually perform the tasks rather than just knowing about them. Since skill testing is such an important aspect of a competency-based vocational training program, observation of the administration and conduct of skills tests was a preferable method of assessing skills testing at Woodland. However, due to the unstructured nature of the individualized programs at Woodland it would have required a great deal of time to conduct any such assessment. Therefore, skill testing was assessed by assessing the skills tests and interviewing instructors and students (See section [Delimitations, page 3]).

To assess the instruments used for performance testing (Woodland calls them "skills tests") the computer generated a random listing of 50 skills tests from the 3500 different
"competencies" taught. Each of the vocational areas concerned was visited and, where possible, two copies of each of the skills tests was obtained. These were rated against nine criteria developed from the literature, particularly from Jones and Whittaker (1975). The criteria for judging the skills tests are:

(1) Are there printed instructions for the candidates?
(2) Do the instructions state exactly what the student is to do, under what conditions, and to what standards?
(3) Are there standardized clear and unambiguous instructions for the test administrator/rater which would contribute to improved inter-rater reliability?
(4) Is there a standardized, clear, and explicit marking scheme such as checklists, rating scales, etc. and, if rating scales are used, are they behaviourally anchored?
(5) Is the emphasis on product and process or only on the product?
(6) Are there well developed test specifications?
(7) Are there multiple forms of the test?
(8) If there are multiple forms, are they really parallel or equivalent forms?
(9) Is the test congruent with the instructional objectives?

In this case also there were two important questions which could not be answered: (1) Does this test measure an important aspect of the job? and (2) Is the test an adequate sample of important tasks or functions? It would have required a content specialist in each trade area from which the sample was drawn to answer these questions.
Assessing the Instructional Objectives

In a competency-based training program data to provide direction to students so they can demonstrate mastery of a competency or ability is provided by instructional objectives and the instructions for the test. These may be supplemented by the instructor but good instructions and objectives should be unambiguous and clear enough to stand on their own. The test instructions were dealt with on the test assessment. Instructional objectives had to be collected for each of the sample of 50 knowledge tests and the sample of 50 performance tests to provide two samples of 50 instructional objectives which were selected as randomly as the tests. These objectives were to be rated against criteria developed from the literature on the writing of instructional (behavioural) objectives.

Although early program development efforts at Woodland started out using objectives written as terminal and enabling objectives, subsequent program development has departed from this format. As a result, only the older unrevised learning modules contain objectives. Similar information is provided in the newer modules but it is now contained on the introductory page and on the pages describing the skills tests. Therefore, it was necessary to collect the information which originally appeared in the objectives and judge it as though it were part of an objective. The objectives were judged on the basis of six criteria derived from the literature, specifically from: Dick and Carey (1990), Gronlund (1982), Mager (1962), and McAslan (1979). The criteria for judging the objectives are:
(1) Is there a goal statement which clearly describes why it is necessary to learn this competency?
(2) Do the behavioural outcomes (or the tests) specify exactly what performance will be required?
(3) Do the behavioural outcomes (or the tests) specify the conditions under which the performance will be required?
(4) Do the behavioural outcomes specify the level of performance (standards) required for each performance?
(5) Is the behavioural outcome congruent with the goal statement?
(6) Is the objective written as an objective? If not, then the data that were available had to be collected from several different pages.

Interviews

All the data referred to above can be characterized as objective data since they were collected systematically and randomly from pre-existing sources according to predetermined procedures. It can be argued that these data were still evaluated subjectively since the criteria used in making the judgments were influenced by availability and some personal considerations (time and money to be expended); however, decisions regarding the data collection are based on the opinions of eminent experts in the field and do not represent the opinions of the rater. The remaining data were collected primarily through interviews with students, instructors, and administrators. A structured interview schedule was utilized to interview a random sample of 43 students from the student population of approximately 730. Because of the very unstructured nature of this institution with full-time and part-time students attending when they choose and the lack of formal
classrooms and classes, it was difficult to select a random sample in the traditional ways. Therefore, a list of random numbers was used to generate a random listing of program areas which were then visited in a random manner. Upon walking into a program area the number of students present were noted and numbered sequentially and then a random sample of two to four students was selected by using another random number sequence. The students interviewed were asked to participate on a one to one basis to which most readily agreed. For the few who declined the next students in the sequence were chosen as replacements. The students were guaranteed anonymity and this was assured by not asking the student’s name. Demographic data such as the students sex, age, racial origin were noted on the interview schedules and the resulting summaries were then compared with population parameters supplied by the school administration. Since the population parameters compared favorably with the sample statistics, the student sample was judged to be reasonably representative of the entire population on these characteristics selected for comparison.

The differences noted in the 21-25 and >40 age groups could be due to the fact that older students are more likely to be part time students and in attendance in the evenings. I did not conduct any interviews at that time of day.
Table 1

Characteristics of Interview Sample and Campus Student Population

<table>
<thead>
<tr>
<th></th>
<th>Sample</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percent</td>
<td>Number</td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>26</td>
<td>60.5%</td>
<td>344</td>
<td>57.7%</td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>17</td>
<td>39.5%</td>
<td>252</td>
<td>42.3%</td>
<td></td>
</tr>
<tr>
<td>Age &lt;20</td>
<td>11</td>
<td>25.1%</td>
<td>149</td>
<td>23.4%</td>
<td></td>
</tr>
<tr>
<td>21-25</td>
<td>14</td>
<td>32.5%</td>
<td>159</td>
<td>25.0%</td>
<td></td>
</tr>
<tr>
<td>26-30</td>
<td>6</td>
<td>14.0%</td>
<td>104</td>
<td>16.4%</td>
<td></td>
</tr>
<tr>
<td>31-35</td>
<td>5</td>
<td>11.6%</td>
<td>71</td>
<td>11.2%</td>
<td></td>
</tr>
<tr>
<td>36-40</td>
<td>5</td>
<td>11.6%</td>
<td>79</td>
<td>12.4%</td>
<td></td>
</tr>
<tr>
<td>&gt;40</td>
<td>2</td>
<td>4.6%</td>
<td>74</td>
<td>11.6%</td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>6</td>
<td>14%</td>
<td>92</td>
<td>15.5%</td>
<td></td>
</tr>
</tbody>
</table>

A random sample of ten instructors was chosen to be interviewed using a somewhat less structured interview schedule than was used with the students. A list of all the instructors in the competency-based programs was obtained from the administration and numbered sequentially. A small program for generating random numbers on a personal computer was used to select the ten instructors. These instructors were then approached and the interview requested. In the two cases where the instructors were unable or unwilling to be interviewed, another instructor from the same program area was selected at random for interview.

A number of very unstructured interviews were conducted with administrators including the principal, vice-principal, deans, program coordinators, and a few other instructors for the purposes of soliciting cooperation and assistance, explaining what I was
doing, and gathering data or clarifying points. These, along with informal observations from having spent many days in the institution, resulted in various impressions particularly as related to the organizational climate. It would have been preferable to have used some more empirical measures of organizational climate such as the study done by Dr. Adrian Blunt at Wascanna Institute in 1985; however, the firing of Mr. Kalyn as principal in September tended to create an environment which was not conducive to the conduct of such an endeavor.
Chapter 4. Analysis of Data

Computer Test Bank Item Analysis

The sample of 242 multiple-choice test items drawn from the computer were analyzed according to the twelve criteria by scoring each item as a one or a zero on each of the twelve criteria on a scoring sheet. If the item met the criteria, it was scored '1', if not it was scored '0'. The results were then entered into a spreadsheet, totalled, and percentages calculated.

Only 119 of the 242 items, or 49.28 percent, met all of the criteria. One hundred and twenty three items (50.72%) had one or more faults capable of reducing the effectiveness of the item.

Table 2

<table>
<thead>
<tr>
<th>Number of Flaws</th>
<th>Number of Items</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>119</td>
<td>49.28</td>
</tr>
<tr>
<td>1</td>
<td>58</td>
<td>24.06</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>15.38</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>8.73</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.69</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>.86</td>
</tr>
</tbody>
</table>

Total number of items 241 100.00

To the uninformed observer this may not seem to be a serious problem; after all, what is one flaw out of eleven possible flaws?
However, if one flaw is enough to permit students to guess the correct answer, then it is enough to render the item useless as a determinant of whether or not the learner has acquired the knowledge being tested. If the flaw does not entirely give away the answer but permits the student to eliminate one or two of four possible answers, then it raises the possibility of getting the answer correct by guessing from 25 percent to 50 or even 75 percent. Ought an item which gives the student a 75 percent chance of guessing the correct answer to be used in a test of competency?

The twelfth criteria used to judge the items was whether the item tested a level of learning above the lowest level of Bloom's taxonomy. Only (14 of 242) six percent of the items could be said to test levels of knowledge higher than simple recognition or recall. Items among the six percent tested comprehension or application levels, the second and third levels in Bloom's taxonomy. No items were identified which test knowledge in the cognitive domain beyond the third level. The six levels of Bloom's cognitive taxonomy are: (1) knowledge, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6) evaluation.

To provide a check on the validity of the rating by the primary rater and to determine whether test item construction training would benefit Woodland's instructors, the sample of items was broken down into sub-samples and given to a group of eleven undergraduate students to rate the items on the same twelve criteria. The students were enrolled in a university credit vocational education class at the University of Saskatchewan concerned with the production and administration of tests including multiple-choice item tests. The students had received instruction from the professor regarding the criteria and had some practice in writing multiple-choice items. Fifteen items were selected from the first thirty of the computer drawn sample. These items were
chosen by the primary rater and included perfectly constructed
items and items with each of the types of errors represented by
the judging criteria. All eleven students rated this sub-sample and
these ratings were compared with the primary raters ratings. The
inter-rater agreement on this common sample averaged 83 percent
(range 79 to 89 percent).

The remaining 210 items were divided into sub-samples of 15
items each and these were randomly assigned to pairs of students
so each sub-sample was rated by the primary rater and two
students. The ratings of the two students and the primary rater
were compared with each other to obtain a percentage of
agreement. This was done by entering all the ratings on
spreadsheets and then summing the spreadsheets one at a time.
Thus, for a given sub-sample there were three spreadsheets, one
for each student and the primary rater. Two of these spreadsheets
were summed by laying one on top of the other. Any cell which
contained two zeroes summed to zero; any cell which contained two
ones summed to two; but any cell which contained a one and a zero
(showing disagreement of the raters) summed to a one. The
number of ones in the resulting spreadsheet divided by the number
of cells multiplied by 100 equals the percentage of disagreement
between the two raters. The percentage of agreement equals 100
minus the percentage of disagreement.

\[
100 - \frac{\text{Number of ones}}{\text{Number of cells}} \times 100 = \text{Percent of agreement}
\]

This process was repeated three times for each sub-sample
which resulted in a total of seventeen calculations of percentage of
agreement. These were entered in another spreadsheet and
averaged to provide a mean percentage of agreement between
raters. It was felt that a relatively high level of agreement
between raters would: (1) permit a high degree of confidence in the
ratings of the primary rater, and (2) demonstrate the potential for
The percentage of agreement between raters averaged 85 percent, (range 73 to 98 percent); so, the twelve raters agreed on their ratings of the items in 85 percent of all possible cases. This is generally considered good inter-rater reliability (Huck, Cormier, & Bounds, 1974).

Analyzing the Computer Tests

The sample of computer generated tests was analyzed in a similar manner to that utilized to analyze the test items. Only 59 percent of the tests (46 of a total of 78) had at least 10 items, the minimum number which most experts consider necessary to establish test reliability. The number of items varied from 3 to 33 with a mean of 10.

Table 3

Computer Test Analysis Summary

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Number Good</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. At least ten questions.</td>
<td>46</td>
<td>59</td>
</tr>
<tr>
<td>2. Items independent of all others</td>
<td>75</td>
<td>96</td>
</tr>
<tr>
<td>3. Items grouped by objective.</td>
<td>78</td>
<td>100</td>
</tr>
<tr>
<td>4. Position of answer varies.</td>
<td>76</td>
<td>97</td>
</tr>
<tr>
<td>5. Alternatives in logical order.</td>
<td>26</td>
<td>33</td>
</tr>
<tr>
<td>6. Length of correct answer varies</td>
<td>72</td>
<td>92</td>
</tr>
<tr>
<td>7. Test higher levels of learning.</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8. Related to learning objective.</td>
<td>78</td>
<td>100</td>
</tr>
</tbody>
</table>

| Total Number | 78 | = 100% |

Ninety-six percent of the tests have items which are entirely independent of all other items; that is, one item could not provide a
clue to the answer for another item. This is not surprising considering the low ratio of items in the test bank to test items drawn. Computer systems personnel at Woodlands stated that there are approximately three items for each one drawn. However, a simple calculation, the total number of items (73,000) divided by the total number of competencies (3500) shows a ratio of 20.86 items per competency. With an average test of 10 items per competency, then the mean ratio of items to items drawn is only two to one. Test experts in this field consider a ratio of ten to one to be the minimum acceptable for a good test bank. (Baker, 1974b)

Therefore, the fact that 96 percent of the items are independent is more likely to be a measure of the simplicity and inadequacy of the test bank than a product of a well planned item selection strategy.

All the test items are grouped by objective in the computer database because they are grouped by competencies with essentially one objective per competency. There is no logical order or order of increasing cognitive level of items because only six percent of all items test higher levels of learning. Thus, as mentioned previously, 94 percent of the items test the lowest cognitive level of learning.

In 97 percent of the tests analyzed the position of the correct answer varies from one item to the next. The computer does a good job of randomly positioning the correct answers, however, it does not have the capability of placing alternatives in a logical order in those cases where such a logical order exists or is desirable. Only 33 percent of all tests examined met this criterion. The relative length of the correct answer varies 92 percent of the time.

All items were judged to be related to the stated learning objective; but, since the raters were not content experts, this rating means that there was nothing that a lay observer might detect as a problem with the congruency between the test items
and the objectives. All the test forms were judged to be equivalent. This observation was not surprising considering that there were so few items for selection at the objective level. If there were a larger proportion of items testing higher levels of learning and if the tests were constructed from more adequate test specifications, then these findings would become much more significant.

The sample of computer tests was divided into sub-samples and rated by the same group of undergraduate vocational education students who rated the computer test bank items. The ratings by these students were compared with the ratings of the primary rater in a similar manner to that used for the item analysis. Spreadsheets were constructed and overlaid to calculate a percentage of rater agreement. All students rated a common sub-sample of computer tests. The percentage of agreement on this common sub-sample ranged from 74 to 96 percent with a mean of 82 percent. In addition, pairs of students rated another sub-sample with their ratings being compared to the primary rater and to each other. The average agreement on these sub-samples was 83 percent (range 69 to 98 percent). Levels of inter-rater reliability greater than 80 percent are generally considered to be good, and it must be remembered that, although these students may have received more training than the average instructor at Woodlands, they have experienced only a level of training in testing and measurement considered to be required for entry level teaching positions in the high school.

Analyzing the Skills Tests

The skills tests were rated on the nine criteria specified in Chapter three. The ratings showed that 46 tests or 100 percent of
the tests in the sample had printed instructions for the candidate, most being printed at the back of each module.

Table 4

Summary of Skills Test Analysis

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Number Meeting</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Printed instructions for student.</td>
<td>45</td>
<td>96</td>
</tr>
<tr>
<td>2. Performance, conditions, standards.</td>
<td>42</td>
<td>89</td>
</tr>
<tr>
<td>3. Instructions for test administrator.</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>4. Checklists, rating scales (BARS)</td>
<td>46</td>
<td>100</td>
</tr>
<tr>
<td>5. Emphasis on product and process.</td>
<td>37</td>
<td>79</td>
</tr>
<tr>
<td>6. Well developed test specifications.</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>7. Multiple forms of the test.</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td>8. Multiple forms are equivalent.</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td>9. Test congruent with objectives.</td>
<td>41</td>
<td>87</td>
</tr>
</tbody>
</table>

Total Number 46 = 100%

In 80 percent of cases there was emphasis on both process and product. Although this emphasis appears on the tests, the instructor interviews revealed that time often does not permit the instructor to monitor the process. The performance required, the conditions, and the standards were stated in 91 percent of cases, however, in many cases the performance required is very briefly described. Only 41 percent of the sample had multiple forms of the test; however, in all cases the multiple forms were judged to be equivalent forms. The tests were rated as congruent with the instructional objectives for 89 percent of the sample. Only one program (9 percent) had any kind of instructions for the test administrators which would contribute to inter-rater reliability, and
the same program area was the only one where there were any kind of test specifications. All tests had a standardized marking scheme and in all cases this was a simple yes/no check list. Much is left to the subjective discretion of the instructor administering the test.

Analyzing the Objectives

As previously mentioned, Woodlands has departed from the use of stated instructional objectives in their modules. Only 25 percent of the objectives judged were actually written as objectives according to the conventions established by Gronlund (1981, 1982), MacAsham (1979), Mager (1962), or Tyler (1949).

Table 5

Summary of Analysis of Instructional Objectives

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Number Meeting</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clear goal statement.</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2. Performance specified.</td>
<td>38</td>
<td>81</td>
</tr>
<tr>
<td>3. Conditions specified.</td>
<td>40</td>
<td>85</td>
</tr>
<tr>
<td>4. Performance level specified.</td>
<td>47</td>
<td>100</td>
</tr>
<tr>
<td>5. Outcome congruent with goal.</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>6. Written as an objective.</td>
<td>12</td>
<td>26</td>
</tr>
</tbody>
</table>

| Total Number                           | 47             | 100%    |

For the rest, the information was gleaned by searching through the modules for an introduction, a page specifying learning activities, and the skills tests. Those which were written as instructional objectives followed the style of Mager (1962) and, therefore, only
eight percent of cases from the sample showed a goal statement which provided a rationale for learning the competency. In six percent of cases the behavioural outcomes were judged to be congruent with the goal statement. Most of the objectives, or objective material, concentrated on the specification of behavioural outcomes which is really only the evidence of achievement of the goal and not the desired outcome. (Gronlund, 1981, 1982; McAshan, 1979)

Eighty-one percent of the objectives clearly specified the performance required and 85 percent specified the conditions for the performance. These conditions are contained in the skills test and are basically just a list of necessary materials. It is a rare case where any other situational variables or restrictions are mentioned. All the objectives specified the performance level required. However, in most cases this was a common statement such as: (1) "acceptable performance requires a score of 80 percent" (for the computer test) and, (2) "acceptable performance requires obtaining a "yes" rating for all criteria on the check list." The check list is always found with the skills test at the back of the module.

In general, except for the goal statement, the information normally contained in an instructional objective can be found if one searched for it, but it does not communicate to the student a clear unambiguous description of what is to be accomplished, to what standards, and why.
Analyzing the Student Interviews

A sample of 43 students was interviewed to determine how they actually use the testing system. Most, 37 (83 percent) of students usually or always study all the material recommended or presented by the study guides before trying the computer tests. The majority 33 (77 percent) felt that it was usually or always necessary to study all the material before trying the test. A Similarly 36 (84 percent) usually and 1 (2 percent) always passed the computer tests on their first attempt.

Fifty six percent said it was never possible to pass a computer test just by guessing, but 44 percent said it was sometimes possible to do so. Most students also admitted it was common knowledge that certain modules are easy to pass without much studying. Ninety-five percent of students agreed that it was impossible to cheat on a computer test while another five percent thought it might sometimes be possible. An on-site examination also indicated that the computer testing area was very secure.

Each study guide has built-in self-tests to help students prepare for and assess their readiness for the computer-generated knowledge tests. Approximately half of the students, (53 percent) always and one third (33 percent) usually complete the self-tests before trying the computer tests. Only 56 percent said that they usually or always use the self tests as an indicator of readiness for the computer tests while 65 percent said they usually or always use the self-tests as an indicator of what they need to study more before trying the computer tests. Students tend not to use the self-tests as a pretest, 77 percent never and 23 percent sometimes. Also, students do not see a great deal of similarity between the self-test items and the computer test items. Ninety-five percent of students say that you cannot pass the computer tests by just
studying the self-tests. Therefore, the self-tests do not serve to short cut the learning process.

Most students, having failed a computer test, either study the same material again (86 percent) and/or go to an instructor for help (65 percent). Only 21 percent usually and 21 percent sometimes look for different material to study. Thirty-five percent say they can usually find alternate materials if they actively seek them. Eighty-four percent find the instructors helpful when they have failed a computer test. Obviously, students take the easy way out, either restudy the same material they didn't learn the first time or go to an instructor. The instructor provides a verbal explanation which is the most common alternate mode of instructional presentation. Instructors appear to spend most of their time in interviews with students. So much so that they may have too little time for improving the testing system.

In one program area, I was told that instructors made a rule that they would not assist students until they had tried the computer test at least twice. The idea was to force the students to take more responsibility for their own learning and not to use the instructors as an easy way out. However, this method was not very successful since students would go to a counsellor and have them review the test.

The vast majority of students (85 percent) use basically the same study methods throughout the course. They also perceive that most other students function in very much the same way as they do. It would seem, therefore, that although the competency-based system is different from other educational and training experiences that students may have encountered previously, they learn a method of coping with it but do not necessarily learn how to become more effective, independent learners. In a traditional setting, the teacher is the primary source of instruction on a group basis and reading materials are a secondary source. In this
setting, reading materials are the primary source of instruction and the instructor becomes the secondary source but on an individual basis.

Most students (74 percent) believe it takes a sufficient level of skill to pass the skills tests and 78 percent feel it would take much more practice to achieve a high level of skill. Although 90 percent feel that improving their competence is a good reason to do the extra practice, they want to do it on the job. Fifty-one percent say that students sometimes help each other on skills tests and 65 percent said instructors sometimes help students pass skills tests by providing extra help during the test.

In summary, students learn how to use the system by studying the material in the learning guides, doing the self-tests, and then trying the computer tests. They do not vary or change their study habits or look for supplementary materials. If they fail a computer test, they will review the same material and/or approach an instructor for help. The computer testing service is very secure and the items are different enough from the self-tests to inhibit attempts to pass simply by studying the self-tests. Students feel that most of them function in much the same manner. They sometimes help each other on skills tests and instructors also provide assistance with skills tests.

**Analyzing the Instructor Interviews**

Ten instructors randomly selected from a population of 68 by a computer program were interviewed and asked a total of 37 questions plus their comments. The first four questions were designed to determine instructor's background experience and qualifications.
Table 6

**Instructors Educational and Experience Background**

<table>
<thead>
<tr>
<th>Instructors</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Apprenticeship</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical School</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University Voc/Tech cert.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Ed.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other university degree</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Work Experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (years)</td>
<td>14</td>
<td>3</td>
<td>0</td>
<td>41</td>
<td>23</td>
<td>15</td>
<td>32</td>
<td>19</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>Supervisory</td>
<td>10</td>
<td>26</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journeymans licence</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Measurement training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University classes</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshops</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total years of experience ranged from 0 to 41 years with a mean of 17.5 years. Years of supervisory experience ranged from 0 to 26 with a mean of 5.8 years. Educational background ranged from one person without high school to two persons with two university degrees each. Three of these people had been to technical school, served apprenticeships, and obtained a vocational-technical teaching certificate. A fourth had been to technical school but worked in an
area that was not licenced in any way. Two people had university
degrees which were directly applicable to the occupational domain
in which they were instructing. Two others had two university
degrees, a B.Ed. and one other degree which did not necessarily
relate directly to the domain in which they were instructing. One
person had been to technical school, obtained a journeymans
licence, and a B.Ed. degree. One person did not have a completed
high school or any other formal qualifications.

Five persons have taken one or more university level classes
in student evaluation. Four indicated they had attended one or
more seminars at the institution but this included two of the above
five. Three persons had not had any training or education in
evaluation. Nine out of ten have attended at least one professional
development seminar but these were not necessarily concerned with
evaluation.

Eight of the ten instructors said they had been involved in
the development of student tests for use at Woodland. When asked
about the steps that were followed in developing these tests,
deciding what should be tested and to what standards, it turned
out that the instructor's personal opinions were the primary
sources. Two indicated that some type of job analysis had been
carried out somewhere else and then utilized as the basis for
program development at Woodland. Four said that provincial
advisory boards provided some guidance.

Paper and pencil tests, hands-on practical testing, either on
the job or as a simulation, were given as the testing alternatives
which had been considered. Some felt that they were locked into
the system implemented at Woodland and had no alternatives. Only
two persons seem to have considered or implemented alternatives,
both had B.Ed. degrees.

When asked how standards for testing were determined, the
instructors opinions again surfaced as the predominate criterion
while the guidelines of provincial advisory boards were also mentioned. Two instructors said that feedback from practitioners in the field was used to set standards.

Woodland commonly uses checklists in the evaluation of the skills tests. Instructors were asked where these originated. Four indicated they developed their own, three got them from books, and five obtained them from someone else. These responses were not too specific but basically indicated that only four out of ten had been involved in developing their own checklists for their evaluations.

The next series of questions was designed to determine what instructors knew about basic concepts of evaluation. They were asked what determines the difficulty of a test. The answers were: time, accuracy, task difficulty, amount of knowledge required, and level of cognitive functioning. Only two persons (both have a B.Ed.) realized that anything other than time or accuracy were applicable criteria. Next the instructors were asked how they would change the level of difficulty of a test. These were open ended questions and the conclusion reached was that only the two instructors with a B.Ed. had any understanding of the concept of test item difficulty.

Instructors indicated that their personal feeling or judgment was the sole basis for deciding whether or not a test discriminates between good and poor performance. There was no empirical basis for their decision. When asked whether they used test specifications, only three claimed to do so although one of the three admitted to only doing it mentally. Essentially most instructors (except the ones with a B.Ed.) did not know or understand what was meant by test specifications or blueprinting. The same was true of the concept of pilot testing of tests. All instructors indicated there was no emphasis within the institution on the pilot testing of new tests. Nine of ten said that there was
no such thing as an institutional manual on the production of tests. One person identified a pilot version of material on testing which had been developed locally by one of the program writers. This material was basically a good summary of literature by Gronlund and Gagne.

The skills tests are rarely ever or never modified. Only those instructors with an educational background had any conception of the need to do so. When changes are made, they are prompted primarily by program changes and secondarily by problems with students or available resources. One instructor said there wasn't any time available to instructors to make changes. Most (7 out of 10) instructors felt that the skills tests are as realistic as can be expected in this type of setting. That does not mean the tests accurately simulate on-the-job performance. Two said their tests were not at all realistic.

In most cases (7 out of 9) the student's tests and score sheets were filed. In two cases they were returned to the students. Six instructors felt that a relatively standard method of test administration was used, but three said there was considerable variation in the way instructors administered skills tests. Six instructors said there was no test manual or guidance available on test administration in the institution. Three thought there was such a manual but could not identify or supply it. Six instructors said there was a standard, objective method of marking employed for skills tests but three said there was not. Clearly there is a large degree of misunderstanding among instructors about what resources are available and little evidence, that if the resources exist, they are being used extensively.

Time permitted to complete skills tests was a factor in five out of nine cases. One instructor said time was not relevant in that field while three others did not time students but admit it is important on the job. Those that do time students allow several
times the standard on-the-job rate. Most (7) instructors said they assess the process which the student follows as well as the product. Two said that it is essential to do so. Two others said that there is not enough time for instructors to assess process; they only look at the product. Do instructors coach or assist students during skills tests? Three admitted to doing so and one said sometimes. Five said they never provide coaching. But 65 percent of the students said that instructors sometimes help students pass skills tests by providing extra help during the test.

Only one of ten instructors could say there was any kind of comprehensive final exam in their program although two claimed that the final modules were comprehensive in that they built on the other modules of the program through proper sequencing. However, not all students graduate from the program. Some leave when they can obtain employment and may never reach those final modules. Three of the programs required a compulsory practicum and one utilized a voluntary practicum. Six programs do not use any kind of practicum.

The next three questions dealt with the computer test bank. When asked about guidelines for creating test items for the computer test bank, four instructors used the advice of program writers, specifically, those developed and presented at a seminar by one of the institutes program writers. Two used guidelines from Gronlund's material, which is essentially what the aforementioned guidelines are. Most instructors were aware that there are materials available on the production of high quality test items. Four instructors said they regularly modify or revise test bank items although this may only be one or two a week. Five instructors rarely or never modify or revise computer test items. What prompts such changes when they do occur? Nine times out of ten it is the result of student complaints. The rest of the time it is the result of program changes.
The last seven questions all related to student attitude. All instructors felt that student attitude was important to future successful employment. Seven of nine felt that it is very important. The most common method of teaching good attitudes is modelling by the instructor, but, five of the instructors had modules or materials built into the program specifically to teach attitudes. Five had no formal process and one did nothing to teach attitudes. This person felt that the instructor did not bear any responsibility for student attitudes. When it comes to deciding on desirable attitudes, eight of ten used their personal opinions. One of the eight was able to make these standards very explicit and spelled them out in detail to the students. The other seven had implicit standards which they were not able to explain. Two mentioned the opinions of advisory board personnel as having a bearing on their attitude standards.

When asked how they decide what behaviours are indicative of a given attitude, most rely on their personal experience. As mentioned previously, one instructor spelled out these expectations very explicitly to students and used observation of specific criteria. One instructor does not assess student attitude. One "doesn't really try to assess student attitude" and the majority who do admit to doing it in a very subjective manner. Except in the two cases where attitudes are not being assessed, instructors say students know they are being assessed. Only in two cases does attitude assessment have anything to do with whether a student passes or fails but it does have an influence on instructor's recommendations of students and thus their future employability.

Everyone stated that attitude assessment is important and most believed they are capable of assessing other peoples attitudes, but few could say how it ought to be done.
Summary of Data Analysis

The analysis of data gathered through a multi-method approach has revealed many areas in which the student assessment system at Woodland Campus could be improved. Data collected from the test item databank show that half of the items are technically defective. Data collected from the computer generated knowledge tests indicates the system of generating knowledge tests functions as a valuable administrative tool but is a simple and unsophisticated system. Data collected on instructional objectives cast doubt on their utility as a means of communicating to students what is to be learned, why, and to what standards. The skills testing is standardized but simplistic. Many instructors are not qualified or skilled enough in their new profession. The next chapter uses the findings from this data analysis to draw conclusions and make recommendations for improving the student assessment system at Woodland Campus.
Chapter 5: Conclusions and Recommendations

Conclusions

This study examined the student evaluation system of S.I.A.S.T. Woodland Campus competency-based vocational training programs. Only programs which were vocationally oriented and competency-based were examined. The study did not attempt to examine any program in detail, but, was concerned with the overall quality of the institution’s student assessment system. Thus, the conclusions and recommendations apply to the competency-based training programs as a whole and not to any one program in particular. There are some programs which fare much better than others from this review and some that fare much worse.

Knowledge Testing

Testing of students prerequisite knowledge, that is, whether the students have learned the necessary theoretical knowledge to progress to skills testing, is done through the computerized testing system at Woodlands. Thus knowledge testing in this situation is primarily concerned with the instrumentation.

Instrumentation - the computer testing system.

Senior administrators of Woodland Campus have publicly claimed that the computer managed learning and testing system at that institution is a state of the art system. (Woodland Campus: An Institute with a Difference) However, the reality appears to be that the testing system is actually a simplistic system, implemented by a contemporary computer system. A data bank of approximately
73,000 test items and a small mainframe computer system (IBM 4381) do not make a system complex or state of the art. The total pool of items is subdivided into smaller pools for each of the approximately 3500 competencies. To generate a test the computer uses a simple stratified random selection procedure to take items from the pool. Originally, all items were multiple-choice with four possible answers, but the system has been upgraded to permit binary-choice and fill-in-the-blank type items. Although this is an enhancement, it is a simple improvement and does not greatly improve the testing system. There is no merit to using different types of items just for the sake of having different types; the different types of items have their own advantages and disadvantages and should be utilized accordingly. Test items for any test ought to be generated according to carefully developed test specifications rather than a simple random selection from a pool. These test specifications will be different for each test and, therefore, the computer programming required to generate such tests would necessarily be much more complex than the present system at Woodland.

The computer is only a tool to reduce the effort required to store and retrieve test items and it ought not to drive the system, but make it possible to do efficiently what would otherwise be difficult or impossible to do on a large scale. This system functions well in one respect, it makes it possible to generate 500 or more different tests in one day, administer, score, and return the results to students and faculty in a matter of hours. The system also performs well in maintaining student progress records. Thus, from an administrative point of view, it is an excellent management tool. From an instructional perspective, the system is mediocre at best. Assessment is always a compromise between the desirable (requiring larger resources) and the practical (requiring
fewer resources), Woodland's system definitely favours the practical.

If Woodland's computer system actually had the 500,000 items in its test bank reported in IBM's sales literature (Woodland Campus: An Institute with a Difference), and if the quality of those items was good, then Woodland's would have the basis for advancing to a more sophisticated testing system. However, Woodland's only has 73,000 items (15% of IBM's claim). The quality of the items is low since 50 percent contain structural faults, which can limit their effectiveness, and only approximately six percent of all items measure higher levels of learning outcomes.

Data - instructions and feedback.

Students do not seem to have any difficulty using the computer testing system, therefore, it can be said that the orientation program and the directions from the computer are sufficient to communicate to students how to use the system. It is not so clear that the instructional objectives or objective materials (conditions, performance required, and standards) actually communicate to students what they need to learn to pass the tests. The objectives do not communicate why it is necessary to learn the knowledge and only in rare cases do they communicate what the person needs to learn or to what level. Having to achieve a score of 80 percent on every test, regardless of difficulty, may be a standard but it is the simplest possible standard to state and, when stated without regard to test difficulty, it is not evidence of a quality testing program.

The computer tells the student whether they have passed or failed. It does not provide any feedback on which items were answered incorrectly and does not provide any guidance for
further study or activities to correct the deficiency. The students go to the instructors for this feedback. Providing feedback on items that were answered incorrectly is an inefficient use of the instructor's time and reduces the time available for other important instructional activities. The self-tests in the modules should serve to help students determine their readiness for the computer tests. However, they do not seem to be effective in doing this for two major reasons: (1) the self-tests are no better developed than the computer tests and thus are of limited usefulness, and (2) the students have not (in spite of the orientation program) learned how to function as independent learners.

**Incentives.**

Students are permitted three attempts to pass a computer test. If they are unsuccessful on all three attempts, they are locked out of the system until an instructor or aide reinstates them for another attempt. Whether the student passes the test on the first, second, or third attempt, or even the fourth, makes little difference to the student, thus, there is little reason other than personal pride to be successful the first time. Originally there was no incentive in terms of time or money to complete a test successfully the first time since the student paid the same amount regardless of how long they took to complete their program. Now the fee structure is based on the program and the length of time, so the student can save some money by completing the program in a shorter time span. Since it requires more time to retake computer tests, it costs more money.
Performance Testing

Proponents of CBL are adamant that the system is superior to traditional programs because students must be able to perform the skills of a job, not just know about them. They also claim greater objectivity because criterion-referenced testing is used. However, Schalock (1981) says, "The Achille's heel of CBL is the quality of assessment."

Instrumentation.

There are basically two problems involved in assessing human competence. The first is the technical problem of measurement of performance and the second is the setting of standards by which competence is to be judged. Although it is never possible to eliminate all subjective judgement, both problems should be solved in a manner which ensures validity, reliability, and as much objectivity as possible. A sensitive system is necessary to reduce the associated problems of false positives and false negatives.

The general consensus of experts in both education and business is that validity begins with comprehensive and rigorous job analyses. The great majority of programs at Woodland have not been derived from job analyses. Although an unvalidated curriculum development model "Metaskills Profiling" which uses one faculty member as a content expert and a program writer was used, nothing has ever been published about this model. Rusnell and Collins (1989) also gave the institution a failing grade on competencies "derived from a rigorous analysis of the occupation" and further noted, "In our opinion, committee meetings and group discussions are not equivalent to rigorous analysis." (p. 10)

Not only competencies, but, also the standards of performance
should be derived from the job analyses. There is no evidence that standard setting at Woodland Campus involved anything more than the instructor's opinions and the organizational directive that the pass mark on knowledge tests be 80 percent. Therefore, because of lack of evidence to the contrary, the validity of the student performance assessment at Woodland is doubtful. There is insufficient evidence to substantiate the claim that the Woodland Campus assessment system results in valid assessment of student performance.

Woodland's performance testing is relatively simplistic and probably no better than the testing done at any traditional post-secondary vocational training institution. All performance rating is done on a simple checklist which indicates pass or fail, acceptable or unacceptable performance. All criteria from the most important to the least, carry the same weight. There are no behaviourally anchored rating scales used and the pass/fail standard is strictly the subjective opinion of the rater. There are no test specifications and most instructors do not appear to know what test specifications are. The same is true for the institution's test manuals. Although some instructors do collaborate in an attempt to ensure the fairness of their ratings, many do not and thus it is probable that inter-rater reliability is low.

Instructors admit the degree of realism (fidelity) of the testing situations is low, but, believe it is as realistic as is practical in any institutional setting. This is not surprising coming from people who are not knowledgeable about measurement and performance evaluation. Essentially, the instructors use the same methods of testing which they experienced when they were students in more traditional institutions over two decades ago. Instructors believe their testing differentiates between competent and incompetent students but none of them can specify how this is so. The concept of "pilot testing" which was recommended by
Rusnell and Collins (1989) has still not been adopted in this institution.

**Data - instructions and feedback.**

The same objectives are used for performance as are used for knowledge goals and therefore the previous comments on the use of objectives also applies here. Some tests have written instructions for the student but many do not. Most depend on the instructor to explain to the student what is required. Explanations are usually not extensive since the test is often a repetition of an exercise the student has completed previously. Feedback to the student is delivered verbally by the instructor in most cases. In those cases where written tests are employed, the instructor writes comments on the tests which are returned to the student.

**Incentives.**

The individualized system of testing at Woodland permits students to be tested whenever they feel ready for the test. If a student fails a test, the sole consequence is that the test has to be repeated. Therefore, there is no necessity to be certain of having mastered a module before challenging the test. In some cases there is little difference between a practice exercise and a performance test. Some instructors have stated that there is essentially no difference. If the student does not do well, then it is just an exercise and the results are not recorded. When the student can do the exercise sufficiently well to satisfy the instructor, then it is recorded that the student has passed the test. This simple testing can be time consuming and does not provide the student with any external incentive to practice alone.
Motivation and Attitude Measurement

The third element of the human factor is motives and is concerned with the individuals motivation and attitudes. It may be the most difficult to assess objectively and is often the most neglected.

Instrumentation, data, and incentives.

Little instrumentation exists at Woodland for the assessment of student’s attitudes and motivation. A few programs have modules to teach attitudes and some make a definite effort to assess student attitude and motivation, but, most do not. The instructions provided to students to teach attitudes and motivation are limited to direct exchanges between students and instructors. Most instructors believe attitudes are an important component of the training process and some attempt to assess student attitudes, usually informally, but only one instructor could state explicitly how it was done. In most programs attitude assessment has no bearing on the successful completion of the program. The only incentive for students to develop positive attitudes and motivation is to obtain a good recommendation from the instructor. This could be more a function of interpersonal relationships than the student’s attitude or motivation towards the knowledge and skills to be learned.

Recommendations

Improve the Organizational Climate

Improving the organizational climate is a necessary prerequisite to improving the student evaluation system. S.I.A.S.T.
needs to adopt a modern management paradigm as promoted by contemporary management authorities such as Peters (1984, 1985, and 1987) and Drucker (1974). This paradigm recognizes many of the fallacies of the older authoritarian management practices including the idea that management is a generic practice and, therefore, a manager can manage any activity. Knowledge of the product or process is not necessary if management is a generic capability. The corporate office of S.I.A.S.T. needs to recognize the necessity of an intimate understanding of both the product to be created and the process by which it is created. Instructors and their supervisors need more than trade qualifications or experience. They also need knowledge and skills in measurement and performance evaluation.

Two other related fallacies which must be addressed are: (1) those who can perform a job can teach it to others, and (2) those who can perform a job are qualified to rate the performance of others on the same or similar jobs. The first fallacy is erroneous because it assumes no special knowledge or skills are required to instruct or teach. It assumes an instructor only needs the knowledge and skills of the trade being taught. This is not so. In fact, the ability to teach frequently requires an entirely different set of knowledge and skills. As evidence of this, in sports some of the best athletes (players) have failed to become good coaches while some of the best coaches were never outstanding players (athletes). The knowledge and skills required to teach a competency are different than the knowledge and skill required to perform the competency. In addition, a good instructor usually needs some management skills, thus, the instructor actually requires knowledge and skills in three different areas: (1) the trade, (2) education and training, and (3) management of human and material resources. Few individuals have all three capabilities to the same high standard.
Some people believe the government thought CBL was such a "teacher-proof" mechanistic system that almost anyone could make it work. Also, since CBL was a "new" methodology, knowledge and skills in "traditional" education were detrimental and would hinder the implementation of the new system. Therefore, an effort was made to hire, whenever possible, instructors with only trade qualifications. If this is so, it would explain some of Woodland's problems. It would also explain why one former administrator (before he knew he would be leaving Woodland) stated that he despaired of ever being able to improve the system because so many instructors could not see any value in obtaining knowledge and skills in education and training.

The second previously mentioned fallacy is that a person who can perform a job will know how to assess others doing the same job. This view is in error for exactly the same reason as the first, it takes different knowledge and skills to assess performance than to do the job. As the literature review indicates, assessing human performance is a field which requires highly specialized knowledge.

The organizational climate at Woodland and S.I.A.S.T. in general has fluctuated over the years since construction started on the original Northern Institute of Technology which later became Woodland, but has been very negative overall. Cochrane (1990) has provided a partial chronology of events under the title "SIAST's Killing Fields" in the critical education newsletter Scrutiny. (1990). This chronology deals with the many staff cuts, changes, and firings during SIAST's brief history. The review does not include the labour problems arising from other factors including what some perceive as a clumsy attempt to destroy the instructor and staff union. The organizational climate improved for a short period following the appointment of Mr. Peter Kalyn as principal on January 7, 1989 but deteriorated again when he was fired on October 1, 1990.
Scrutiny conducted a survey of three CBE institutions including Woodland in May 1991, the results published in the Sept/Oct 1991 issue, focused on CBE rather than the organizational climate; however, the low return rate (21.2 percent) and some of the comments may be indicative of a poor organizational climate. Although I guaranteed anonymity to all those I interviewed for this study there were still some who did not wish to risk their positions by participating in the study. There appears to be a definite lack of trust and respect between the staff of Woodland and the S.I.A.S.T. corporate office. This may be attributable to the managerial practices of S.I.A.S.T. and the Government of Saskatchewan. It is not likely that significant improvements in the student evaluation system at Woodland can be achieved until the organizational climate is improved.

Knowledgeable, dedicated teachers are recognized to be essential to a successful CBL program (Gamson, 1979). S.I.A.S.T.'s management needs to realize that dedication and commitment must be earned by the organization, they are not subject to command.

Improve the Instrumentation for Instructors

Some proponents of CBL promote the idea that because instructors of individualized instruction are not required to prepare for or deliver group instruction, they have the time required to tutor individual students (Blank, 1982). Indeed, some see this as the primary function of the instructor and claim that the ratio of students to instructors can be increased two or three fold. However, tutoring students is only one function for which instructors are responsible. Even in institutions where the quality of the primary instructional materials is excellent and students are able and willing to learn on their own, it is still necessary for someone other than the instructor to carry out the other
professional duties required such as program design and improvement, instructional design, the design and construction of evaluation instruments, the administration of tests, and all the other administrative duties associated with implementing CBE.

Most of these professional activities cannot be accomplished in the few minutes available between student consultations. At Woodland Campus instructors share open offices which have glass doors and walls. The offices are adjacent to shop areas and are not suitable for intellectual activities requiring concentrated thinking. If instructors are to be involved in duties other than tutoring and counselling students, then they need facilities and time suitable to the accomplishment of these other duties.

**Improve the Incentives for Instructors**

Gamson (1979) emphasizes the need for committed professionals to successfully implement good CBL programs. She points out, as does Peters (1984, 1985, and 1987), that commitment comes from ownership. People look after that which they own. The best programs observed by the writer at Woodland were those in which instructors with the necessary expertise had invested themselves in the program. These people have used their talents and many hours of their own time in program development. Unfortunately, the institution does not appear to provide any recognition or reward for this extra effort. Senior administrators at Woodland need to develop ways of promoting program ownership among staff.

**Improve the Data to Instructors**

Woodland's staff need a performance assessment system. It is in their own interest to have one. Just as students need a feedback system to inform them about their progress and what they
need to improve, instructors and administrators also need such a system. The purpose of such a system should be the improvement of professional performance rather than the rating or ranking of personnel for advancement or promotional purposes. Gilbert (1979) believes that the area with the greatest potential for improving human performance is in the provision of information that can be used as feedback to people so they know how well they are performing.

As in many workplaces, there will be much opposition to such an idea, especially because of the poor organizational climate within SIAST and the instability and lack of employment security. Thus, improvement in the organizational climate is a necessary prerequisite for such a system. The assessment system should be developed with the participation of as many people as possible. There is no reason for this system to be limited to a top-down assessment. A two-way assessment system could be more effective and beneficial (Gauthier, 1987; Jacobs, 1986; Schweir, 1982).

**Improve the Human Element**

Competency-based vocational training has great potential for improving the effectiveness and efficiency of student learning, but, it is not a simple mechanistic system. It’s successful implementation requires professional instructors, knowledgeable and skilled in the teaching-learning process. Instructors need to recognize that when they became instructors they changed professions, not just jobs. Their new profession requires the acquisition of new knowledge and skills and the instructor must bear the primary responsibility for obtaining them. Persons becoming instructors take on new responsibilities, not just to the organization, but also to the students. The adoption of a professional ethic - responsibility to the client, the community, and society - is part of the new
profession. It is up to the instructor to obtain the education and training necessary to fulfill these new responsibilities to the highest level possible.

The organization has a responsibility to assist by arranging for such training, and should offer incentives and rewards for those who seek and apply new knowledge and skills, but, the organization cannot learn for the individual instructor. All instructors should be lifelong learners. How are students to be encouraged to be lifelong learners if the instructors are not lifelong learners themselves?

An instructor's role is more than just demonstrating and explaining job skills. Dr. Gerald Sankey, now retired from the University of Saskatchewan, has said that more people lose jobs because of poor attitudes than because of lack of skills. Learning appropriate attitudes is therefore as important as learning skills. (Rusnell and Collins, 1989) Teaching attitudes is primarily achieved through modelling and instructors are the role models. How can an instructor teach students the value of good performance if they choose not to demonstrate superior performance themselves?

Motivation is a very personal attribute. An organization may establish incentives and rewards but the best it can do is to create an environment in which people may be motivated to perform. No one can directly motivate another person. Motivation is an internal condition under the control of the individual. Wlodkowski (1988) maintains that one of the most important skills an instructor can teach students is to help them learn how to motivate themselves. If an instructor is to teach students about self-motivation, then surely the instructor must be capable of working, to some extent, as a self-motivated individual. In the final analysis, instructors must develop the motivation to perform as professionals.
Improving the Student Assessment System

The purpose of this study of Woodland’s student evaluation system was to test the model by using it to evaluate Woodland’s student evaluation system. It was suggested that the logical application of the model would result in recommendations for the improvement of the student evaluation system. It might seem that these particular recommendations should have been placed at the beginning of this chapter instead of the end; however, the reason for not doing so is that the previous recommendations are necessary prerequisites for organizational development.

Improve the computer testing system.

(1) Examine all items in the test bank and rewrite the 50 percent which have technical flaws.

(2) Increase the number of items by adding items many of which are required to test higher levels of knowledge.

(3) Improve the system by implementing a new item selection design based on comprehensive test specifications. Ideally, this system would also provide better feedback to students so that they would not necessarily need an instructor’s assistance if they fail a test. The feedback should warn students of areas in which they are weak and what to do for further study. This would require upgrading the existing instructional design to provide alternate methods of instruction (Blank, 1982; Guskey, 1985; Rusnell and Collins, 1989).
Improve the performance testing system.

(1) Substitute planned flexibility for the existing ad-hoc flexibility. Whenever possible students should be tested in small groups instead of individually. Instructors administering performance tests should not be interrupted for other purposes. This would make more efficient use of instructors time and improve the reliability and validity of the testing.

(2) Tests and testing standards should be developed from a comprehensive occupational analysis and not just from one or two instructors opinions. Tests should also be pilot tested with successful practitioners to establish concurrent validity.

(3) Improve the fidelity of testing - the testing situation should be realistic - as close to an on-the-job situation as possible. Use practicums for this purpose but be sure the raters are trained in rating student performance. This will improve both the reliability and validity of the performance testing.

(4) Introduce more comprehensive testing which tests student performance across functions instead of only on isolated tasks. Since competence involves performance across functions, more comprehensive testing is necessary to ensure the validity of testing and the subsequent certification of students as competent practitioners.

(5) Operate a comprehensive final examination in a manner similar to an assessment centre. This could be done on a periodic basis and use trained examiners from other institutions and even from industry. It could be open to others (students from other
institutions and non-students) who wish to gain the certification without having attended this institution. In some cases this might even be integrated with the apprenticeship system since there is an existing interprovincial system of challenge exams for journeyman certification.

**Improve the assessment of student attitudes.**

(1) Students should be made aware early in the program, why certain attitudes are important and highly desirable, the specific attitudes expected of them, the behaviours which are recognized as indicative of these attitudes, how they will be assessed, and the consequences of this assessment. Peters (1984, 1985, 1987) points out that only that which is measured gets done, and that it is essential to reward the type of behaviours desired.
Chapter 6: Summary

The Study Goal

The goal of this study was to develop a model of excellence in performance evaluation and demonstrate its application in one practical situation. The seven element performance model was developed from an extensive literature review. It borrows heavily from the work of Gilbert (1978, 1982a, 1982b) but expands Gilbert's model into a more comprehensive model by adding one element, the organizational climate, and clarifying the interactional relationships between all of the elements. Literature on performance measurement, evaluation, and certification from education, commerce, and other disciplines provided strong theoretical support for the proposed model.

The Study Objective

The objective of the study was to compare the S.I.A.S.T. Woodland campus student evaluation system with criteria suggested by the performance assessment model which serves as an exemplar of the present state of the art of performance evaluation. Since proponents of competency-based vocational training are adamant that CBE produces people who can "actually perform" as opposed to just "knowing about" an occupation, the ability to measure performance accurately is essential to a competency-based vocational training program. As Peters (1984, 1985, 1987) repeatedly points out, we can only create and control the quality of that which we can adequately measure.
Assessing Human Performance

Assessing human performance is itself a performance and a student evaluation system is a performance system and can be analyzed as such. The literature on performance measurement and evaluation is related to the seven element performance model to indicate the conditions which facilitate good performance assessments. These conditions are briefly summarized below.

The Environment

The environment within which the performance assessment takes place must provide the assessor with the necessary data, instrumentation, incentives, and climate.

Data.

The environment must provide the assessor with information on the purposes, methods, and importance of the performance assessments. It should also provide feedback on the quality of the assessments being completed and how they could be improved.

Instrumentation.

The assessment instruments should be developed so that they contribute to valid, reliable, objective, and sensitive assessments which are also practical. The instrument should be accepted by both assessors and assessees as valid for their purposes.
Incentives.

Performance assessments should be used to accomplish their stated purpose and must be seen to be used for that purpose. The motivation to conduct assessments is seriously effected by the belief that they are not used. The quality of performance assessments should be assessed and the assessors rewarded for producing quality assessments.

Climate.

The psychological climate of the organization sets the tone within which performance assessments are made. Whether this climate is authoritarian, persuasive, consultative, or democratic will have much to do with the manner of conduct and type of performance assessments which are carried out. It will also be a major influence on the purpose of the assessments.

The Personal Factor

Knowledge.

The individual assessor needs to have knowledge of:

1. the job/performance being assessed,
2. the instrument(s) being used in the assessment,
3. personal and statistical biases and their consequences,
4. measurement theory and practice, and
5. benchmarks and standards for the performance.
Skills.

The individual assessor should have skills in the ability to:
(1) communicate and interact with people,
(2) observe performance accurately,
(3) record observed performance accurately,
(4) judge observed and recorded performance,
(5) assess limitations imposed by the environment,
(6) assess ratees ability to compensate for the environment,
(7) apply measurement theory and avoid statistical biases,
(8) recognize and avoid personal biases, and
(9) provide an accurate and usable report of the assessment.

Motives.

The individual assessor must believe in the importance and utility of the assessments being compiled and the importance of making accurate and bias-free assessments. If the assessor does not accept the value and utility of the performance assessment, then the quality of the assessment is likely to be less than it could be. The assessor must realize that he or she is part of the performance environment of the person being assessed and, therefore, has an influence (positive or negative) on the performance of the person being assessed.

Application of the Model

The seven element model was applied to Woodland’s student evaluation system as a practical example of its utility. This required an evaluation of the processes used to assess the three elements of a student's personal factor. The four elements of the
environmental factor were assessed in relation to their influence on the student evaluations. Instructors were assessed as part of the environment which interacts with the students. The environment within which the instructors perform was assessed in relation to its influence on the instructors.

It can be seen that for each level of the organization, the assessor or supervisor is part of the performance environment of the person being assessed. Thus, a complete evaluation of any performance system has multiple levels with each level interacting with the levels above and below it. Most performance assessments are done from the top down; however, since each level has a direct effect on the level below it as well as above it, perhaps an effective assessment system should also run from the bottom up.

**Methodology**

This study used a multi-method approach because no one method of data collection could provide all the data required. Whenever possible hard data were collected from the campus. The computer item-bank was systematically sampled to obtain a sample of test items and two samples of equivalent forms of the computer generated tests were obtained. The computer was used to generate a listing of a random sample of skills tests which were then obtained from the respective departments. All of this data were analyzed by rating it on objective criteria derived from the literature. Students selected at random were interviewed to obtain data on how students use and view the system. Instructors were also selected at random and interviewed to determine their knowledge, skills, and attitudes. Longer, semi-structured interviews were held with a number of administrators to obtain other data.
Analysis of Data

The Computer Testing System

The computer testing system includes a relatively large data bank of approximately 73,000 test items, 95 percent of which are multiple choice items. Analysis of the sample of items indicates that the average quality of these items is poor with approximately 50 percent having technical faults which reduce items utility. The large number of items tends to obscure the fact that the ratio of items to items drawn per test is only 2:1 on average. Simple random selection of poor quality items from an inadequate pool of items does not result in the production of high quality tests.

The strength of computer utilization is that it makes it possible to generate large numbers of individual tests on demand and keep track of the results; however, the value of these tests is doubtful. Some instructors dismiss the computer testing as merely a formality required by the system and not an asset to either students or instructors. These instructors believe that their own performance tests provide the only measurement of any value. The simplicity of the computer testing is such that even newly trained instructors should be able to construct better quality tests than the computer can presently generate.

A small group of undergraduate vocational education students was used to rate sub-samples of the items and computer tests. Although these students have only introductory training in testing and measurement, the percentage of agreement on their ratings was adequate (average over 80 percent) to indicate that instructors with even this small amount of formal training could improve the quality of items in the test bank. Improving the overall quality of the computer testing would require a much better test
specificational design procedure which could then be transformed into a computer program.

**Performance Testing**

The quality of the performance testing at Woodland varies considerably from one program area to another. In general, there are no test specifications or guidance for test administrators and few program areas with multiple or parallel forms of the performance tests. The skills tests follow a standardized format which utilizes checklists as a marking scheme and specifies that the student must receive a "yes" rating on each criteria on the checklist. The format is very simplistic with the outward appearance of objectivity but is actually very subjective with most decisions being left entirely to the discretion of instructors, many of whom know little about testing, measurement, or performance assessment.

**Objectives**

The concept of instructional objectives is probably not well understood within the institution. As with the testing system, a general format is followed which like a good formula is supposed to ensure a proper result if followed correctly. However, without an understanding of the formula, what is actually created is a format which is not necessarily useful for the purpose it should serve. Thus, even the 25 percent of objectives which are written appropriately fail to inform the student exactly what needs to be learned, why, and to what level of performance. This is said to be a shortcoming of traditional training programs which CBL is supposed to overcome.
Student Interviews

Analysis of the student interviews reveals that the computer testing system is very secure and well understood by students but does not provide them with useful feedback, only the information that they have passed or failed. Most students do not utilize the available resources to the best advantage nor do they learn how to learn. Instead, they only learn how to cope with a different type of learning environment. Students indicate that coaching on skills tests is fairly common by other students and by instructors. Most instructors will not admit to this directly but do admit they are not able to closely supervise students engaged in skills tests.

Instructor Interviews

Probably the greatest weakness of instructors is a lack of knowledge and skill in professional educational activities such as measurement and evaluation, and many do not even see the need for this knowledge and these skills. Professional instructors require knowledge and skills in education, as well as in their trade area. The ability to execute the duties and responsibilities of a trade does not automatically endow one with the ability to teach the trade to others or to assess the performance of others.

Conclusions and Recommendations

Competency-based vocational training is a systematic method of delivering instruction which includes many components, each of which must be developed and refined to contribute to the overall effectiveness of the instructional design. Of these components, the measurement and evaluation component may be the most critical. As
Schalock (1981) says, it is the "Achilles heel" of CBL. This component must embody the five essential characteristics of a good assessment system: 1) validity, 2) reliability, 3) objectivity, 4) sensitivity, and 5) functionality. In most cases Woodland's assessment of students is only functional. The student evaluations at Woodland are probably no better, if as good, as those of most traditional institutions.

Recommendations for improving the student evaluation system at Woodland were made in three stages with the accomplishment of each stage being a prerequisite to the accomplishment of the next. This does not mean that none of the recommendations of the third stage, which dealt most directly with the student evaluation system, can be implemented until all the preceding recommendations are accomplished. Rather, it implies that a level of excellence in student evaluation is unlikely to be achieved until the improvements recommended in the first two stages are realized. Thus, improving the organizational climate, the first stage, is a necessary prerequisite to significant improvement in the rest of the system, but the implementation of improvements in the second and third stages could begin immediately.

The second stage of recommendations focus upon the need to improve the knowledge and skills of instructors, provide better facilities and scheduling so instructors can properly execute professional duties, and to provide incentives for instructors to perform those duties at a high level of competence.

The third stage of recommendations dealt with direct improvements to the student evaluation system. Recommendations were made to improve the quality of test items in the data bank and expand the size of this data bank by 75 to 80 percent. A more sophisticated computer testing design was recommended to improve the quality of measurement and the usefulness of the testing by providing students with better feedback and direction.
Recommendations were made to improve the quality of the performance testing through: better planning and less ad-hoc flexibility; the development of valid standards; the pilot testing of tests; and more comprehensive and realistic testing to ensure competence across functions.

A greater effort should be made to affect students attitudes positively so they may also develop professional attitudes towards themselves and their careers. This may be the most difficult of all the recommendations to realize. It may also be the most important.

The conclusions reached and the recommendations made for the improvement of the student evaluation system at Woodland were based on the different types of data gathered at the institution. The quality of the vocational training that Woodland provides varies considerably from one program to another. Thus, the number and type of recommendations should not be taken as an indication that the campus overall does a poor job of training its students, but rather the effectiveness of the campus to evaluate the training outcomes falls far short of the institutions claims and the standards that the students pay for and ought to receive under CBVT.
REFERENCES


Blunt, A. (1986a). Determining the Validity and Reliability of Criterion Referenced Test Scores. (Unpublished manuscript located at Program Development Division, S.I.A.S.T. Wascana Campus, Regina, Saskatchewan)


Appendix A

Box 1113,
Prince Albert, Sask.
S6V 5S7
Sept. 9, 1989

Dr. F. Barry Brown
EdCCV College of Education,
University of Saskatchewan,
Saskatoon, Sask. S7N 0W0

Dear Barry,

The thesis I propose to do is based on the question of whether or not the competency-based format for vocational education as implemented at SIAST, Woodland Campus in Prince Albert, actually measures competence as it purports to do, and, if so how well it does. Study has convinced me that competence is a judgement of a measure of human performance and that there are basically three components or factors that are directly involved in competent performance. These are: knowledge, skills (psychomotor, cognitive, and affective), and attitude.

Woodland Campus utilizes a large computer data base of nearly 250,000 questions to generate multiple choice tests for the measurement of prerequisite knowledge. This is followed by practical training exercises and then performance tests. I propose to draw a sample of questions from the data base and analyze them according to criteria derived from the literature on essential qualities such questions should meet. I would also draw a sample of complete tests from the data base and analyze these against criteria for the construction of complete multiple choice tests. A similar procedure would be followed with the performance tests except that it will also be necessary to interview and observe instructors in their administration of these tests.
Enclosed is a copy of page 73 of the report by Dr. J. Collins and Dr. D. Russnel on their recent evaluation of Woodland Campus. These evaluators recommended that it would be timely to examine, revise, and upgrade the existing student evaluation system; thus, I believe my study would be of value to SIAST at this time. One of the recommendations (3.) is to conduct an item analysis of each question in the data base. I would consider doing such an analysis on the questions drawn for my sample after the completion of my initial analysis, however, at the present time I have some reservations about the usefulness of such a procedure.

In order to complete this study I will need the approval of the SIAST board, the cooperation of the staff at Woodlands, access to pertinent documents, and access to the data stored in the computer. Mr. Peter Kalyn has expressed verbal approval of the general idea but requires a written request. Therefore, would you please arrange to have the necessary correspondence forwarded to them at your earliest convenience.

Thankyou.

Yours truly,

T. B. Dent
October 6, 1989

Mr. Peter Kalyn
Principal
SIAST, Woodland Campus
Prince Albert, SK.

Dear Mr. Kalyn:

Mr. Tom B. Dent has approached you in respect to carrying out research in your institution for his thesis for the Masters Degree in the Department of Communications, Continuing and Vocational Education.

I have reviewed the study Tom plans to carry out in the area of competence. I feel his proposed study will be a contribution to this area and I am sure, based on Tom's track record with us, it will be of value to your institution.

This will confirm that, as Tom's supervisor, I feel the study has merit and I forward to you a formal request that he be accorded the necessary permission and approval to carry out his study at the Woodland Campus.

If you or your Board have further questions, please do not hesitate to contact me for additional information or classification of Tom's Research interest.

F. Barry Brown
Professor, EDCCV

FBB/sc
Appendix C

Mr. P. Kalyn Principal,  
Woodland Campus SIAST  
Prince Albert, Sask.  
Oct. 12, 1989

Dear Mr. Kalyn,

Re our telephone conversation of October 12, 1989.

The thesis I propose to do is based on an assessment of the student evaluation system utilized in your competency-based vocational education programs. Study has convinced me that competence is a judgement of a measure of human performance and that there are basically three components or factors that are directly involved in competent performance. These are: knowledge, skills (psychomotor, cognitive, and affective), and attitude.

Woodland Campus utilizes a large computer data base of nearly 250,000 questions to generate multiple choice tests for the measurement of prerequisite knowledge. This is followed by practical training exercises and then performance tests. I propose to draw a sample of questions from the data base and analyze them according to criteria derived from the literature on essential qualities such questions should meet. I would also draw a sample of complete tests from the data base and analyze these against criteria for the construction of complete multiple choice tests. A similar procedure would be followed with the performance tests except that it will also be necessary to interview and observe instructors in their administration of these tests.

On page 73 of the report by Dr. J. Collins and Dr. D. Rusnell
on their recent evaluation of Woodland Campus, these evaluators recommended that it would be timely to examine, revise, and upgrade the existing student evaluation system; thus, I believe my study would be of value to SIAST at this time. One of the recommendations (3.) is to conduct an item analysis of each question in the data base. I would consider doing such an analysis on the questions drawn for my sample after the completion of my initial analysis, however, at the present time I have some reservations about the usefulness of such a procedure.

In order to complete this study I will need your approval, the cooperation of the staff at Woodlands, access to pertinent documents, and access to the data stored in the computer. My timetable for this project is to have the proposal completed and accepted by the end of this year, collect and analyze the data by Easter of 1990, and have the thesis completed by September of 1990. Although the data collection will take place early next year, I will need to interview some of your staff in the near future so that I can make the proposal as specific as possible.

Thankyou.

Yours truly,

T. B. Dent
November 1, 1989

Professor F. B. Brown
Department of Communications
College of Education
University of Saskatchewan
SASKATOON, SK
S7N 0WO

Dear Barry:

Re: Tom Dent's Research Project

We would be very pleased to accommodate Tom's request to carry out research in our institution. We see this project as being of benefit to all parties.

As indicated to Tom, I have asked him to work out the details of accessing our data with Rick Sullivan and Larry Fladager of our staff.

Sincerely,

Peter Kalyn
Principal

PK/wv
cc: Larry Fladager
    Rick Sullivan
    Tom Dent
Appendix E

Interview Schedule for Student Interviews

Objective: To determine the process actually used by students in proceeding through the learning guides and evaluation exercises. Do students carefully study the material for each module, do the self tests, use the results to achieve remediation, and then proceed to the computer test and finally the skills tests? What kind of remediation do they seek? Do they redo the self-tests and do they use them as an indicator of readiness for the computer test? If students fail the computer test what course of action do they follow? Do they restudy the same material, look for alternative materials, seek further explanations from the instructor, or just try the test again? Do students try to short cut the system and if so how often are they successful? What do students think others do? Do students strive for a "pass" or for a higher level of skill? What level of skill do students think a "pass" represents?
Interview Schedule for Student Interviews

1. What program are you enrolled in?

2. How long have you been in this program?

3. Do you carefully study all the material in the study guide and recommended by the study guide before trying the computer test? N S U A

4. Do you complete the self-tests in each module before trying the computer tests? N S U A

5. Do you use the results of the self-tests to tell you if you are really ready for the computer test? N S U A

6. Do you use the self-tests as a pretest; that is, you go directly to the self-test and try it and if you pass it you try the computer test without studying the module? N S U A

7. Do you use the results of the self-tests to tell you what you need to study more before you do the computer test? N S U A

8. Do you find that the self-test questions are very similar to or the same as the computer test questions? N S U A

9. Is it necessary to carefully study all the material in the study guide before trying the computer test? N S U A

10. Is it possible to pass the computer test just by guessing? N S U A

11. Can you pass the computer tests by just studying the self-tests and the self-test answers? N S U A

12. Do you pass the computer tests on the first try? N S U A

13. If you fail a computer test do you restudy the same material presented in the study guide? N S U A
14. If you fail a computer test do you look for additional and different material to study? \(\text{N S U A}\)

15. If you fail a computer test do you go to an instructor and ask them to explain the material to you? \(\text{N S U A}\)

16. If you fail a computer test do you just try it again without further study? \(\text{N S U A}\)

17. If you fail a computer test, can you find additional and different material to study? \(\text{N S U A}\)

18. If you fail a computer test, are the instructors helpful in providing explanations of the things you did not know? \(\text{N S U A}\)

19. Do you use the same procedure now that you did when you started or have your study methods changed? \(\text{N S U A}\)

20. Do other students you know carefully study all the materials in the study guides before trying the computer tests? \(\text{N S U A}\)

21. Are other students you know successful at passing the computer tests without carefully studying all the material in the study guides? \(\text{N S U A}\)

22. Is it possible to get someone else to do a computer test for you? \(\text{N S U A}\)

23. Is it common knowledge among students that there are certain modules that are easy to pass without studying the material? No Yes Unc

24. What level of skill do you think it takes to pass the skills tests in your field? Low Suff High

25. Would it take much more study and practice to achieve a high level of skill? No Yes Unc

26. Is there any good reason to do the extra study and practice? No Yes Unc
27. Do students help each other on the skills tests?  

28. Do instructors help students pass skills tests by providing hints, extra guidance, advice, etc.?  

Comments by student:

Comments of Interviewer:
Appendix F

Interview schedule for interviewing Instructors

Objective: To determine:

a) how tests and test items are developed and modified
b) level of expertise of instructors in measurement
c) what efforts are made to affect student attitudes
d) how student attitudes are assessed and if this assessment actually makes any difference to students passing or failing.
e) how skills testing is actually carried out
Interview Schedule for Interviewing Instructors

1. What trade experience and qualifications do you have? licences? years? supervisory?

2. What educational background do you have?
   high school?
   apprenticeship?
   technical school?
   university (non-degree)
   university degree (education)
   university degree (other)

3. Have you had any specific training in the measurement and evaluation of student performance?
   university classes?
   workshops?
   self-study?

4. Have you attended any of the "in house" seminars put on for the professional development of instructors here at Woodlands?
   Yes ( ) No ( )

5. Were you involved in the development of any of the student tests utilized here at Woodlands?
   Yes ( ) No ( ) Skip to #10

6. What steps did you follow in deciding what knowledge, skills, and attitudes students should be tested for?

7. What alternatives did you consider in deciding how students should be tested?
8. How are the standards for your tests determined?

9. If you use checklists in your skills tests, where did they originate? Yourself? A book? Someone else?

10. What determines the difficulty of a test?

11. If you thought a test was too difficult or too easy, what would you do to change it?

12. How do you know that a test really discriminates between good and poor performance?

13. Do you create and use test specifications from which to design your skills tests?
   Yes ( )  No ( )

14. Do you do any pilot testing of a new test?
   Yes ( )  No ( )  Sometimes ( )
15. Is there any such thing as a testing manual which establishes methods of testing and rating students?
   Yes ( )  No ( )  Don't know ( )

16. How often do you modify your skills tests?

17. What prompts you to modify a skills test?

18. What happens to the data sheets or student test sheets after the skills test is completed?

19. Are the testing conditions for the skills tests realistic, that is, a close simulation of on-the-job working conditions?

20. Do all raters (instructors) utilize a standard method of test administration, that is, they all conduct the testing in the same way?
   Yes ( )  No ( )  Don't know ( )

21. Are there any written instructions or specifications, or a test manual which clearly specifies how skills tests should be conducted?
   Yes ( )  No ( )  Don't know ( )
22. Do all raters (instructors) use a standard and objective method of marking the students process and product?

Yes ( ) No ( ) Don't know ( )

23. Are the students timed or limited in time to complete skills tests?

Yes ( ) No ( ) Don't know ( )

24. Do the instructors supervise the students closely enough during skills tests to be able to assess process as well as just the finished product?

Yes ( ) No ( ) Don't know ( )

25. Do the instructors coach the students during the skills tests?

Yes ( ) No ( ) Don't know ( )

26. Is there any such thing as a comprehensive final test that students must pass before they receive a certificate for passing this course?

Yes ( ) No ( ) Uncertain ( )

27. Is there any kind of compulsory practicum that must be completed before being certified as having passed this course?

Yes ( ) No ( ) Uncertain ( )
28. What guidelines do you use in creating new items for the computer test bank?

29. How often do you modify computer test bank items?

30. What prompts you to modify a computer test bank item?

31. How important is student attitude to success in your field?

32. What do you do in your program to teach good attitudes?

33. How do you decide what attitudes are desirable?
34. How do you decide what behaviors are indicative of a given attitude?

35. How do you assess a student's attitude?

36. Do the students know that you are assessing their attitudes?
   Yes (  )  No (  )  Maybe (  )  Don't know (  )

37. Does your assessment of a student's attitude affect a student's evaluation (passing or failing)?
   Yes (  )  No (  )
   If Yes, How?

Comments: